Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Avaliação Experimental e Comparação de Algoritmos de Seleção de Atributos Importantes com o Algoritmo FDimBF Baseado na Dimensão Fractal

Huei Diana Lee
Maria Carolina Monard
Richardson Floriani Voltolini
Feng Chung Wu

N° 264

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
Agosto/2005

Avaliação Experimental e Comparação de Algoritmos de Seleção de Atributos Importantes com o Algoritmo FDimBF Baseado na Dimensão Fractal*

Huei Diana Lee†
Maria Carolina Monard†
Richardson Floriani Voltolini†
Feng Chung Wu†

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil
e-mail: {huei, mcmonard, rfv}@icmc.sc.usp.br, wufc@unioeste.br

Resumo. Em aprendizado de máquina, a tarefa de pré-processamento do conjunto de dados inclui selecionar os atributos mais importantes para realizar o aprendizado. A seleção de atributos é de fundamental importância pois, no caso de aprendizado supervisionado, atributos não relevantes ou redundantes podem reduzir a precisão e a compreensibilidade das hipóteses induzidas por esses algoritmos. Vários algoritmos para a seleção de atributos relevantes têm sido propostos na literatura. Entretanto, tem sido observado que somente o critério de relevância não é suficiente para a seleção de atributos importantes. Trabalhos recentes têm mostrado que também deve-se levar em conta o critério de redundância para selecionar os atributos importantes, pois atributos redundantes afetam a qualidade das hipóteses induzidas. Vários modelos têm sido propostos para tratar tanto relevância quanto redundância de atributos, porém, alguns desses modelos apresentam um custo computacional muito alto. Um modelo mais recente sugere realizar o tratamento de relevância e redundância como dois processos separados. A vantagem desse modelo é que, por meio dessa separação, é possível diminuir o custo computacional da busca pelo subconjunto que aproxima o subconjunto ótimo de atributos. Neste trabalho é proposto um algoritmo baseado nesse modelo, *i.e.* que separa as análises de relevância e de redundância. Nesse algoritmo encontram-se implementadas duas medidas para realizar a análise de relevância, uma medida baseada em ganho de informação e outra baseada em distância. Quanto à redundância, é proposto o uso da Dimensão Fractal do subconjunto de atributos relevantes selecionados na etapa anterior. Resultados experimentais utilizando vários conjuntos de dados e diversos algoritmos que selecionam atributos importantes, mostram que a Dimensão Fractal é um critério apropriado para filtrar atributos redundantes no aprendizado supervisionado.

Palavras-Chave: Seleção de Atributos, Aprendizado de Máquina, Dimensão Fractal.

Agosto 2005

*Trabalho desenvolvido com o apoio do Instituto de Tecnologia em Automação e Informática – ITAI, do Parque Tecnológico de Itaipu – PTI e da FAPESP Processo nº .04/04885-8.

†Pesquisador do LABI - Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná

Este documento foi preparado com o formatador de textos LaTeX. O sistema de citações de referências bibliográficas utiliza o padrão *Chicago* do sistema bibTeX.

Sumário

Sumário i

Lista de Figuras iii

Lista de Tabelas vii

Lista de Abreviaturas e Variáveis ix

1 Introdução 1

2 Fractais 3

3 Dimensão Fractal de um Conjunto de Dados 4

4 Algoritmo Proposto 5

5 Descrição dos Conjuntos de Dados 7

6 Algoritmos Utilizados 9

7 Configuração dos Experimentos 13

8 Resultados e Discussão 15

8.1 Dimensão Fractal e Comportamento Fractal dos Conjuntos de Dados . . . 16

8.2 Subconjuntos de Atributos Selecionados 20

8.3 Formatos Aproximados de Distribuição dos Valores dos Atributos em Relação aos Atributos Selecionados pelo Algoritmo FDimBF 22

8.4 Performance dos Algoritmos em Relação à Precisão e a Quantidade de Atributos Selecionados 25

8.5 Análise da Significância Estatística dos Resultados 28

8.6 Características dos Conjuntos de Dados Associadas à Utilização da Dimensão Fractal como uma Medida Adequada para a Seleção de Atributos . . . 31

8.6.1 Características Gerais dos Conjuntos de Dados e Adequação do Uso dos Algoritmos FDimBF 33

8.6.2 Padrões Encontrados na Aplicação dos Algoritmos FDimBF para os Conjuntos de Dados 35

9 Considerações Finais 36

Referências 37

A Dimensão Fractal e Comportamento Fractal dos Conjuntos de Dados 41

B Subconjuntos de Atributos Selecionados 53

C Distribuição dos Valores dos Atributos dos Conjuntos de Dados 61

D Performance dos Algoritmos em Relação à Precisão e a Quantidade de Atributos Selecionados 87

E Regras Induzidas com o Conjunto de Dados Meta1 93

F Regras Induzidas com o Conjunto de Dados Meta2 95

Lista de Figuras

2.1 Triângulo de Sierpinsky . . . 3
3.2 Construção do Triângulo de Sierpinsky . . . 4
4.1 Modelo para seleção de atributos (Yu and Liu, 2004) . . . 5
7.1 Configuração dos experimentos . . . 14
8.1 Gráfico gerado utilizando o método *Box Count Plot* — Hungarian . . . 17
8.2 Gráfico gerado utilizando o método *Box Count Plot* — Waveform . . . 18
8.3 Número de atributos selecionados e a respectiva percentagem *versus* o algoritmo de SA . . . 23
8.4 Tipos de formatos aproximados das distribuições dos valores dos atributos 24
8.5 Relação entre percentagem de atributos selecionados, média do erro e erro padrão dos modelos construídos: (a) Modelo geral e (b) Conjunto de dados Pima . . . 26
A.1 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Breast Cancer . . . 41
A.2 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Bupa 41
A.3 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - German . . . 42
A.4 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Hungarian . . . 42
A.5 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Ionosphere . . . 43
A.6 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Pima 43
A.7 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Satimage . . . 44
A.8 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Segment . . . 44
A.9 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Sonar 45
A.10 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Vehicle . . . 45
A.11 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Waveform . . . 46
A.12 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Breast Cancer . . . 46
A.13 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Bupa 47
A.14 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - German . . . 47
A.15 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Hungarian . . . 48

A.16 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Ionosphere . . . 48
A.17 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Pima 49
A.18 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Satimage . . . 49
A.19 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Segment . . . 50
A.20 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Sonar 50
A.21 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Vehicle . . . 51
A.22 Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Waveform . . . 51
C.23 Distribuições dos valores dos atributos – Breast Cancer . . . 61
C.24 Distribuições dos valores dos atributos – Bupa . . . 62
C.25 Distribuições dos valores dos atributos – German – A . . . 63
C.26 Distribuições dos valores dos atributos – German – B . . . 64
C.27 Distribuições dos valores dos atributos – German – C . . . 65
C.28 Distribuições dos valores dos atributos – Hungarian . . . 66
C.29 Distribuições dos valores dos atributos – Ionosphere – A . . . 67
C.30 Distribuições dos valores dos atributos – Ionosphere – B . . . 68
C.31 Distribuições dos valores dos atributos – Ionosphere – C . . . 69
C.32 Distribuições dos valores dos atributos – Pima . . . 70
C.33 Distribuições dos valores dos atributos – Satimage – A . . . 71
C.34 Distribuições dos valores dos atributos – Satimage – B . . . 72
C.35 Distribuições dos valores dos atributos – Satimage – C . . . 73
C.36 Distribuições dos valores dos atributos – Satimage – D . . . 74
C.37 Distribuições dos valores dos atributos – Segment – A . . . 75
C.38 Distribuições dos valores dos atributos – Segment – B . . . 76
C.39 Distribuições dos valores dos atributos – Sonar – A . . . 77
C.40 Distribuições dos valores dos atributos – Sonar – B . . . 78
C.41 Distribuições dos valores dos atributos – Sonar – C . . . 79
C.42 Distribuições dos valores dos atributos – Sonar – D . . . 80
C.43 Distribuições dos valores dos atributos – Sonar – E . . . 81
C.44 Distribuições dos valores dos atributos – Sonar – F . . . 82
C.45 Distribuições dos valores dos atributos – Vehicle – A . . . 83
C.46 Distribuições dos valores dos atributos – Vehicle – B . . . 84
C.47 Distribuições dos valores dos atributos – Waveform – A . . . 85
C.48 Distribuições dos valores dos atributos – Waveform – B . . . 86
D.49 Gráfico Percentagem × Erro - Breast Cancer . . . 87
D.50 Gráfico percentagem × erro - Bupa . . . 87

D.51 Gráfico percentagem × erro - German 88
D.52 Gráfico percentagem × erro - Hungarian 88
D.53 Gráfico percentagem × erro - Ionosphere 89
D.54 Gráfico percentagem × erro - Pima 89
D.55 Gráfico percentagem × erro - Satimage 90
D.56 Gráfico percentagem × erro - Segment 90
D.57 Gráfico percentagem × erro - Sonar 91
D.58 Gráfico percentagem × erro - Vehicle 91
D.59 Gráfico percentagem × erro - Waveform 92

Lista de Tabelas

1.1 Formato padrão do conjunto de exemplos 1
5.1 Resumo dos conjuntos de dados 9
6.1 Características dos algoritmos de SA 13
8.1 Informações associadas à dimensão fractal dos conjuntos de dados 19
8.2 Resultado da análise dos gráficos de comportamento dos conjuntos de dados quanto à característica fractal 20
8.3 Resumo da quantidade de atributos selecionados por cada um dos algoritmos e suas respectivas percentagens 21
8.4 Formatos da distribuição aproximada dos valores dos atributos 24
8.5 Média de erro e erro padrão para cada conjunto de dados e algoritmo considerado 25
8.6 Algoritmos presentes nos gráficos 27
8.7 Classificação dos algoritmos em relação a percentagem de atributos selecionados × erro do modelo construído 28
8.8 Siglas para os conjuntos de dados 29
8.9 Comparação entre o número original de atributos e o número de atributos selecionados pelos algoritmos de SA. Comparação entre médias de erros dos modelos construídos (em negrito resultados estatisticamente significativos) 29
8.10 Comparação entre os números de atributos selecionados pelos algoritmos de SA. Comparação entre as médias de erros dos modelos construídos (em negrito resultados estatisticamente significativos) 30
8.11 Resumo do número de vezes em que cada algoritmo seleciona um subconjunto menor de atributos 31
8.12 Descrição dos atributos da metabase 32
8.13 Resumo das metabases 32
B.1 Atributos Selecionados - Breast Cancer 53
B.2 Atributos Selecionados – Bupa 53
B.9 Atributos Selecionados – Sonar 54
B.3 Atributos Selecionados – German 55
B.4 Atributos Selecionados – Hungarian 55
B.5 Atributos Selecionados – Ionosphere 56
B.6 Atributos Selecionados – Pima 57
B.7 Atributos Selecionados – Satimage 58
B.8 Atributos Selecionados – Segment 59
B.10 Atributos Selecionados – Vehicle 59
B.11 Atributos Selecionados – Waveform 60

Lista de Abreviaturas, Algoritmos e Variáveis

Abreviaturas

AM . . . . . Aprendizado de Máquina
DF . . . . . Dimensão Fractal
D_2 . . . . . Dimensão Fractal de Correlação D_2
DLE . . . . . *Discover Learning Environment*
DOL . . . . . *Discover Object Library*
DSX . . . . . *Discover Dataset Sintax*
IA . . . . . Inteligência Artificial
KDD . . . . . *Knowledge Discovery on Databases*
MD . . . . . Mineração de Dados
MDE . . . . . *Measure Distance Exponent*
MDL . . . . . *Minimum Description Lenght*
pD . . . . . Dimensão Fractal Parcial
SA . . . . . Seleção de Atributos
SU . . . . . *Symmetrical Uncertainty*
TDIDT . . . . . *Top Down Induction of Decision Trees*

Algoritmos

$\mathcal{C}4.5$ . . . . . Algoritmo para indução de árvores de decisão $\mathcal{C}4.5$
CBF . . . . . *Consistency-based Filter*
CFS . . . . . *Correlation-based Feature Selection*
FCBF . . . . . *Fast Correlation-based Filter*
FDimBF . . . . . *Fractal Dimension-Based Filter*
FDimBF(1) . . . . . *Fractal Dimension-Based Filter - medida de informação*
FDimBF(2) . . . . . *Fractal Dimension-Based Filter - medida de ditância*
ReliefF . . . . . Algoritmo para SA ReliefF

Variáveis

M . . . . . Número de atributos
N . . . . . Número de exemplos
m . . . . . Parâmetro do algoritmo ReliefF

1 Introdução

A entrada para um algoritmo de aprendizado supervisionado consiste usualmente de um conjunto de N exemplos (ou casos) de treinamento $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ rotulados com os valores y de uma função f desconhecida $y = f(\mathbf{x})$, onde os valores $\mathbf{x}_i$ são vetores da forma $< x_{i1}, x_{i2}, ..., x_{iM} >$ cujos componentes são valores discretos ou contínuos relacionados aos *atributos* $A = \{A_1, A_2, ...A_M\}$. Ou seja, x_{ij} denota o valor do atributo A_j do exemplo i. Dado esse conjunto de exemplos de treinamento, o algoritmo induz uma hipótese $\mathbf{h}$ que deve aproximar a verdadeira função f, tal que dados os valores $\mathbf{x}$ de um novo exemplo, $\mathbf{h}$ prediz o valor y correspondente. No caso dos valores y pertencerem a um conjunto discreto de N_{C_l} classes, *i.e.* $y \in \{C_1, ..., C_{N_{C_l}}\}$, a tarefa de aprendizado é chamada de *classificação*, tratada neste trabalho. Por outro lado, se $y \in \Re$, essa tarefa é denominada de *regressão*.

O formato atributo-valor, comumente usado para representar exemplos, é apresentado na Tabela 1.1. Como mencionado, em aprendizado supervisionado cada exemplo é associado a uma classe (rótulo), que pode ser discreta (classificação) ou contínua (regressão). No caso de aprendizado não supervisionado, não há a informação sobre a classe associada a cada exemplo.

	Atributos				
Exemplos	X_1	X_2	...	X_M	Classe (Y)
E_1	x_{11}	x_{12}	...	x_{1M}	y_1
E_2	x_{21}	x_{22}	...	x_{2M}	y_2
E_3	x_{31}	x_{32}	...	x_{3M}	y_3
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
E_N	x_{N1}	x_{N2}	...	x_{NM}	y_N

Tabela 1.1: Formato padrão do conjunto de exemplos

O aprendizado de um classificador $\mathbf{h}$ é determinado pelo valor dos atributos. Ainda que teoricamente o uso de um maior número de atributos para descrever os exemplos deveria fornecer um maior poder de discriminação para aproximar f, isso pode não ocorrer, especialmente na presença de atributos irrelevantes e/ou redundantes, os quais, freqüentemente, confundem o algoritmo de aprendizado. Assim, a Seleção de Atributos — SA — é uma área de pesquisa há tempo explorada não somente em estatística mas também em Aprendizado de Máquina — AM — e Mineração de Dados — MD (Liu and Motoda, 1998). Os resultados obtidos, tanto teórica quanto experimentalmente, mostram que a SA melhora a predição de classificadores e reduz a complexidade do modelo $\mathbf{h}$.

A meta da SA pode ser formalizada do seguinte modo (Yu and Liu, 2004): seja $A' \subset A$ um subconjunto de atributos de A, e $f'(\mathbf{x}')$ os valores associados aos vetores correspondentes a A'. O objetivo da SA consiste em selecionar o subconjunto mínimo de atributos A' tal que

$$\mathbf{P}(C|y = f'(\mathbf{x}')) \approx \mathbf{P}(C|y = f(\mathbf{x}))$$

onde $\mathbf{P}(C|y = f'(\mathbf{x}'))$ e $\mathbf{P}(C|y = f(\mathbf{x}))$ são as distribuições de probabilidades das N_{C_l} possíveis classes dados os valores dos atributos de A' e A respectivamente. Esse subconjunto mínimo A' é denominado subconjunto *ótimo* de atributos.

O seguinte exemplo, citado freqüentemente na literatura, ilustra esse conceito: considerando o conjunto $A = \{A_1, A_2, A_3, A_4, A_5\}$ de atributos e $y = f(A1, A2)$ uma função booleana, há somente oito possíveis exemplos tal que $A_2 = \overline{A}_3$ e $A_4 = \overline{A}_5$. Assim, para determinar o conceito meta tem-se: A_1 é indispensável; A_2 ou A_3, mas não ambos, podem ser ignorados já que $y = f(A1, \overline{A}_3)$; A_4 e A_5 podem ser ignorados. Nesse caso, existem dois subconjuntos A' ótimos, $\{A_1, A_2\}$ e $\{A_1, A_3\}$, e a meta da SA é encontrar pelo menos um desses subconjuntos. Entretanto, o número de subconjuntos de atributos cresce exponencialmente com o número de atributos em A e encontrar o subconjunto ótimo de atributos pode ser NP (Kohavi and John, 1997).

Os diversos modelos de SA propostos na literatura podem ser categorizados nos modelos *wrapper* e filtro. O primeiro utiliza o próprio algoritmo de aprendizado para determinar, em cada iteração, a precisão do classificador $\mathbf{h}$ induzido utilizando o subconjunto de atributos selecionados nessa iteração. Ao final é considerado o melhor subconjunto de atributos aquele que melhora a precisão de $\mathbf{h}$ (Kohavi and John, 1997). A maior desvantagem dos métodos *wrapper* é que, além de serem específicos ao algoritmo considerado, são computacionalmente muito caros para conjuntos de exemplos descritos por um grande número de atributos. Diferentemente, o modelo filtro separa a SA do algoritmo de aprendizado que utilizará o subconjunto selecionado. A idéia é filtrar atributos irrelevantes, segundo algum critério, antes do aprendizado ocorrer.

Além dos atributos irrelevantes, tem sido observado que atributos *redundantes* também afetam a precisão dos classificadores induzidos e, portanto, deveriam ser eliminados (Koller and Sahami, 1996; Hall, 2000). Considera-se que dois atributos são redundantes entre si quando seus valores estão correlacionados, parcial ou completamente, tais como os atributos A_2 e A_3 do exemplo previamente apresentado.

Em geral, os métodos de SA selecionam os atributos pela *avaliação individual* ou pela *avaliação de subconjuntos* de atributos. No caso de avaliação individual, os atributos são ordenados considerando a sua importância na discriminação das N_{C_l} classes. Esses métodos somente removem atributos irrelevantes pois espera-se que atributos redundantes tenham a mesma importância na discriminação das classes. Contudo, métodos que avaliam subconjuntos de atributos procurando por subconjuntos mínimos podem remover tanto atributos irrelevantes quanto redundantes. Assim, a maioria dos métodos existentes para a SA que tratam tanto relevância quanto redundância de atributos, o fazem de maneira implícita por meio da avaliação de subconjuntos de atributos. Ainda que esses métodos geralmente apresentem melhores resultados que os métodos que não lidam com a redundância de atributos, seu alto custo computacional pode torná-los ineficientes para conjuntos de dados com alta dimensionalidade. Recentemente foi proposto o uso da abordagem filtro considerando o modelo de tratamento da relevância e da redundância

de atributos como dois procedimentos separados (Yu and Liu, 2004). A vantagem desse modelo sobre o modelo anterior é que, por meio da separação da análise de relevância e de redundância, permite encontrar um subconjunto que aproxima o subconjunto ótimo mas evita o custo computacional do modelo tradicional da busca por subconjuntos.

Neste trabalho investigamos o modelo proposto por Yu and Liu (2004) e propomos o uso da Dimensão Fractal — DF — como procedimento para tratar a redundância de atributos. Ainda que o conceito de DF seja freqüentemente utilizado na detecção de agrupamento de dados e na indexação de estruturas de alta dimensionalidade, não é de nosso conhecimento que a DF tenha sido utilizada para realizar SA para algoritmos de aprendizado de máquina supervisionados, como proposto neste trabalho. Resultados experimentais obtidos com diversos conjuntos de dados utilizando diferentes procedimentos para tratar relevância de atributos e o algoritmo por nós proposto, que usa a DF para tratar redundância, são apresentados. Esses resultados mostram que a DF é um critério apropriado para tratar redundância de atributos.

Este trabalho está organizado do seguinte modo: nas Seções 2 e 3 são apresentados brevemente conceitos sobre fractais e dimensão fractal. Na Seção 4 é descrito o algoritmo proposto neste trabalho. Nas Seções 5 e 6 são descritos os conjuntos de dados e algoritmos de seleção de atributos utilizados neste trabalho. A configuração dos experimentos é descrita na Seção 7. Resultados e discussão dos experimentos realizados são apresentados na Seção 8 e considerações finais são apresentadas na Seção 9.

2 Fractais

Fractais são definidos pela propriedade de auto-similaridade, ou seja, apresentam, parcial ou integralmente, as mesmas características para diferentes variações na escala em que estão sendo analisados. Assim, partes do fractal, o qual pode ser uma estrutura, um objeto ou um conjunto de dados, são similares, exata ou estatisticamente, ao fractal como um todo. Fractais possuem, em geral, características incomuns, por exemplo, o conhecido Triângulo de Sierpinsky — Figura 2.1. Ele não pode ser considerado um objeto Euclidiano unidimensional, pois possui perímetro infinito, nem tão pouco um objeto Euclidiano bidimensional já que possui área nula. Dessa maneira, pode-se considerar uma dimensão fracionária, denominada de Dimensão Fractal (Mandelbrot, 1985).

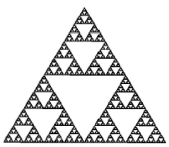


Figura 2.1: Triângulo de Sierpinsky

Muitos dos conjuntos de dados reais comportam-se como fractais. Desse modo, torna-se natural a idéia de aplicar conceitos da teoria dos fractais para a análise desses conjuntos

de dados (Faloutsos and Kamel, 1994).

3 Dimensão Fractal de um Conjunto de Dados

A utilização do conceito de Dimensão Fractal está associada à idéia da existência de redundância nos conjuntos de dados e da possibilidade desses conjuntos serem bem aproximados em dimensões menores. A idéia principal é empregar a DF do conjunto de dados, a qual é relativamente não afetada por atributos redundantes, para determinar a quantidade e quais são os atributos não redundantes segundo o critério de DF (Sousa et al., 2002).

Pode-se definir, desse modo, as idéias de *dimensão imersa* e *dimensão intrínseca*. A primeira idéia corresponde à dimensão do espaço de endereçamento, ou seja, o número de atributos do conjunto de dados. Porém, o conjunto de dados pode estar representando um objeto que possui uma dimensão menor que a do espaço em que está imerso. Assim, a dimensão intrínseca é a dimensão espacial do objeto representado pelo conjunto de dados. Conceitualmente, se um conjunto de dados possui todas as suas variáveis (atributos) independentes umas das outras, então sua dimensão intrínseca será igual a sua dimensão imersa. Porém, toda vez que existir uma correlação entre duas ou mais variáveis, a dimensão intrínseca do conjunto de dados é reduzida de acordo. Usualmente, correlações entre os atributos ou a própria existência dessas correlações não é conhecida. Por meio da dimensão intrínseca do conjunto de dados é possível decidir quantos atributos são necessários para caracterizá-lo. Diferentes tipos de correlação podem reduzir a dimensão intrínseca em diferentes proporções, até mesmo em proporções fracionárias. Desse modo, pode-se utilizar o conceito de Dimensão Fractal como sendo a dimensão intrínseca do conjunto de dados (Traina et al., 2000).

Existem diversas medidas para a DF. Para fractais exatamente auto-similares, *i.e.* que podem ser caracterizados por meio de regras de construção bem definidas, a Dimensão Fractal é dada pela Equação 1:

$$D = \frac{log(R)}{log(\frac{1}{e})} \tag{1}$$

onde R representa a quantidade de réplicas e $\frac{1}{e}$ em que escala as réplicas são geradas a cada iteração.

Para o exemplo do triângulo de Sierpinsky mencionado na Seção 2, a DF seria $D = log(3)/log(2) = 1,58496$, pois são geradas três réplicas em escala $1{:}\frac{1}{2}$ a cada iteração — Figura 3.2.

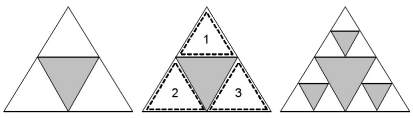


Figura 3.2: Construção do Triângulo de Sierpinsky

Para fractais estatisticamente auto-similares, como conjuntos de dados reais, uma das maneiras para a definição da DF é representada pela Dimensão Fractal de Correlação D_2, que pode ser calculada pelo método *Box Count Plot* (Faloutsos and Kamel, 1994). A idéia consiste, primeiramente, na construção de um reticulado sobre o conjunto de dados de células de lado r. Então, conta-se o número de pontos dentro da i-ésima célula de tamanho r, denominado $C_{r,i}$. A Dimensão Fractal de Correlação D_2 é definida pela Equação 2:

$$D_2 = \frac{\partial log(\sum_i {C_{r,i}}^2)}{\partial log(r)}, r \in [r_{min}, r_{max}] \quad (2)$$

Em teoria, fractais exatamente auto-similares são infinitos. Na prática, conjuntos de dados reais, os quais possuem um número finito de pontos, são considerados fractais estatisticamente auto-similares para um determinado intervalo de escalas $r \in [r_{min}, r_{max}]$ se obedecem um regra de construção bem definida nesse intervalo. Desse modo, a dimensão intrínseca de um determinado conjunto de dados pode ser medida como o coeficiente angular da reta que melhor se ajusta ao trecho linear do gráfico em escala logarítmica de $\sum_i {C_{r,i}}^2$ por r (Traina et al., 2000). Neste trabalho, o termo Dimensão Fractal de Correlação será simplesmente denominado de Dimensão Fractal.

4 Algoritmo Proposto

O algoritmo proposto neste trabalho para a seleção de atributos (Lee et al., 2005,a,b; Lee and Monard, 2003), denominado de *Fractal Dimension-Based Filter* — FDimBF —, pertence à abordagem filtro e segue o modelo proposto por Yu and Liu (2004), ilustrado na Figura 4.1, o qual realiza a seleção de atributos em duas etapas: primeiramente é feita a *análise de relevância* para determinar o subconjunto de atributos relevantes em relação à classe, removendo os atributos irrelevantes; na segunda etapa, por meio da *análise de redundância*, são determinados e removidos os atributos redundantes a partir do subconjunto que contém apenas os atributos relevantes, produzindo o subconjunto final de atributos selecionados.

O algoritmo de Yu and Liu (2004), *Fast Correlation-Based Filter* — FCBF —, utiliza a medida *Symmetrical Uncertainty* (Press et al., 1992) como a medida de correlação para aproximar tanto a análise de relevância quanto a análise de redundância. O FCBF apresenta a vantagem, sobre as abordagens tradicionais para avaliação de subconjuntos de atributos, de que por meio da separação das tarefas de análise de relevância e redundância, ele ameniza o alto custo da busca por subconjuntos de atributos.

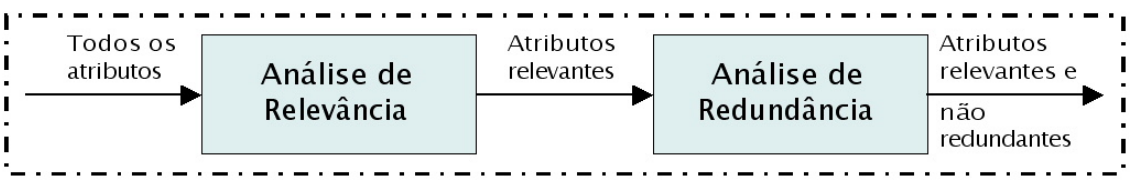


Figura 4.1: Modelo para seleção de atributos (Yu and Liu, 2004)

No algoritmo FDimBF, propomos o uso da Dimensão Fractal como medida para tratar a redundância de atributos. Para realizar a análise de relevância, propomos o uso de duas medidas: uma baseada em ganho de informação, algoritmo FDimBF(1), e outra baseada em distância, algoritmo FDimBF(2). Especificamente, para realizar a análise de relevância em relação ao atributo classe usando a medida de ganho de informação, utilizamos o algoritmo $\mathcal{C}4.5$ (Quinlan, 1993) e os atributos são classificados de acordo com o número de vezes que aparecem nas regras induzidas. Para medir a relevância dos atributos em relação à classe usando uma medida de distância, utilizamos o algoritmo ReliefF (Robnik-Sikonja and Kononenko, 2003) para ordenar os atributos. Esse algoritmo procura pelos exemplos mais próximos da mesma classe e de classes diferentes, e atribui pesos aos atributos de acordo com quão bem eles diferenciam esses exemplos. Esse processo é repetido m vezes. Em geral, m é definido em função do número de exemplos presentes no conjunto de dados.

Como mencionado anteriormente, para tratar a análise de redundância, neste trabalho propomos a utilização da DF. Para medir a DF dos conjuntos de dados, foi utilizada a ferramenta *Measure Distance Exponent* — MDE (Traina et al., 2003). Atributos redundantes, considerando a Dimensão Fractal, podem ser definidos como aqueles que quando excluídos do conjunto de dados não causam uma modificação no valor da DF recalculada. O método usado pelo MDE consiste na medição do valor da DF, D, a partir do conjunto de dados original e do valor da Dimensão Fractal Parcial, pD, ignorando um atributo por vez. Em outras palavras, a pD é calculada tomando-se em consideração todos os atributos exceto o i-ésimo atributo sob observação. O processo continua selecionando o atributo que permite a diferença mínima entre D e pD. Se a diferença é menor que um limiar mínimo, o qual determina quão preciso o conjunto de dados, descrito por apenas os atributos selecionados, precisa ser para preservar as características do conjunto de dados original, esse atributo pode ser considerado como de contribuição pequena para a caracterização do conjunto de dados original. Esse processo continua, considerando o restante dos atributos e fazendo com que $D = pD$ aplicando o procedimento descrito, até que não existam mais atributos a serem removidos. Ao final do processo, os atributos estarão inversamente ordenados de acordo com sua contribuição, em termos de redundância, para a medição da DF do conjunto de dados (Traina et al., 2000).

É interessante ressaltar que muitos dos algoritmos de SA tratam, internamente, apenas atributos nominais. Assim, se o conjunto de dados contém atributos numéricos, eles são discretizados pelo algoritmo antes de efetivamente realizar a SA. Esse é o caso dos algoritmos utilizados neste trabalho. Por outro lado, o algoritmo por nós proposto trata efetivamente atributos numéricos, *i.e.* sem a necessidade que eles sejam discretizados, mas atributos nominais são tratados somente durante a análise de relevância — Figura 4.1 — pois a DF, utilizada para tratar a redundância de atributos, exige que os mesmos sejam numéricos.

5 Descrição dos Conjuntos de Dados

Os conjuntos de dados utilizados para a realização dos experimentos apresentados a seguir, foram selecionados a partir de uma extensa pesquisa bibliográfica de trabalhos publicados na área de seleção de atributos, os quais são freqüentemente referenciados pela comunidade. Nesses trabalhos são utilizados conjuntos de dados reais, naturais e artificiais, sendo:

- reais: extraídos diretamente de bases de dados, por exemplo, de empresas ou hospitais;
- naturais: obtidos de repositório de dados como o repositório da UCI (Newman et al., 1998) e
- artificiais: gerados computacionalmente a partir da função verdadeira $f(\mathbf{x})$ a ser aprendida — Seção 1 na página 1.

A partir dessa pesquisa bibliográfica, foram selecionados 21 trabalhos que utilizam um total de 99 conjuntos de dados diferentes. Esses conjuntos de dados foram ordenados considerando o número de trabalhos nos quais foram utilizados. Após, foram considerados para seleção posterior somente os conjuntos de dados referenciados em pelo menos dois trabalhos. No final desse processo foram selecionados 11 conjuntos de dados supervisionados pouco desbalanceados com atributos numéricos. Esses conjuntos de dados pouco desbalanceados foram selecionados com o objetivo de não introduzir interferências associadas ao uso de um ou outro método para tratar esse problema (Batista et al., 2004).

Todos os 11 conjuntos de dados selecionados, brevemente descritos a seguir, constituem conjuntos de dados naturais obtidos do Repositório de Dados UCI (Newman et al., 1998).

Breast Cancer: o problema é predizer se uma amostra de tecido de mama obtida de uma paciente é maligna ou benigna baseada em dados histológicos;

Bupa: o problema é predizer se um paciente, do sexo masculino, possui ou não disfunção hepática tomando-se como base diversos exames sanguíneos e a quantidade de álcool consumida;

German: nesse conjunto de dados, parte do projeto europeu StatLog (Michie et al., 1994), o problema é classificar pessoas, descritas por atributos como propósito do empréstimo e histórico de crédito, como sendo boas ou más pagadoras, isto é, apresentando risco de crédito bom ou ruim. Esse conjunto de dados é disponibilizado em dois formatos: um contendo somente atributos simbólicos e outro contendo todos os atributos numéricos.

Hungarian: o problema consiste em predizer se um paciente possui ou não doença cardíaca baseado em dados laboratoriais, clínicos e de eletrocardiograma.

Ionosphere: esse conjunto de dados descreve dados sobre radares. Bons resultados desses radares são considerados se mostram evidência de algum tipo de estrutura na ionosfera, caso contrário são considerados ruins os resultados.

Pima: o problema é predizer se uma paciente, mulher de descendência indígena Pima com idade mínima de 21 anos, seria classificada como diabética, segundo o critério estabelecido pela Organização Mundial de Saúde, fornecidos dados clínicos e laboratoriais;

Satimage: esse conjunto de dados, também parte do projeto StatLog, consiste de valores multi-espectrais de pixels de vizinhança 3x3 em uma imagem de satélite e a classificação associada ao pixel central de cada vizinhança. O objetivo é predizer essa classificação, dados os valores multi-espectrais.

Segment: esse conjunto de dados apresenta dados sobre segmentação de imagens. Os exemplos, regiões de tamanho 3x3, foram gerados a partir de imagens de anúncios publicitários. Essas imagens foram segmentadas manualmente para criar a classificação para cada pixel.

Sonar: o objetivo é classificar sinais de sonares refletidos de um cilindro de metal ou de um cilindro aproximado de rocha.

Vehicle: o objetivo é classificar tipos de veículos, usando um conjunto de atributos extraídos a partir de suas silhuetas. O veículo pode ser visto de diversos ângulos. Esse conjunto de dados também faz parte do projeto StatLog.

Waveform: esse conjunto de dados está relacionado com a classificação de tipos de ondas.

A Tabela 5.1 mostra um resumo das características desses 11 conjuntos de dados organizado do seguinte modo:

- # Exemplos: número de exemplos do conjunto de dados;
- # Atributos (num.,nom.): número total de atributos juntamente com o número de atributos numéricos (num.) e nominais (nom.);
- Classes e Classe %: valores e distribuição das classes;
- Erro da CM: erro cometido no caso de novos exemplos serem classificados como sendo pertencentes à classe majoritária — CM e
- ?: existência ou não de valores desconhecidos.

Conjunto de Dados	# Exemplos	# Atributos (num.,nom.)	Classes	Classe %	Erro da CM	?
Breast Cancer	699	9 (9,0)	2 4	65,52% 34,48%	34,48% sobre 2	Sim
Bupa	345	6 (6,0)	1 2	42,03% 57,97%	42,03% sobre 2	Não
German	1000	24 (24,0)	1 2	70,00% 30,00%	30,00% sobre 1	Não
Hungarian	294	13 (13,0)	0 1	63,95% 36,05%	36,05% sobre 0	Sim
Ionosphere	351	34 (34,0)	0 1	64,10% 35,90%	35,90% sobre 0	Não
Pima	769	8 (8,0)	0 1	65,02% 34,98%	34,98% sobre 0	Não
Satimage	4435	36 (36,0)	1 2 3 4 5 7	24,20% 10,80% 21,70% 09,40% 10,60% 23,40%	75,80% sobre 1	Não
Segment	2310	19 (19,0)	1 2 3 4 5 6 7	14,30% 14,30% 14,30% 14,30% 14,30% 14,30% 14,30%	85,70% sobre qualquer atributo	Não
Sonar	208	60 (60,0)	0 1	46,60% 53,40%	46,60% sobre 1	Não
Vehicle	846	18 (18,0)	1 2 3 4	25,10% 25,70% 25,80% 23,50%	74,20% sobre 3	Não
Waveform	5000	21 (21,0)	0 1 2	33,10% 32,90% 33,90%	66,10% sobre 2	Não

Tabela 5.1: Resumo dos conjuntos de dados

6 Algoritmos Utilizados

Os experimentos apresentados neste trabalho foram realizados utilizando quatro algoritmos freqüentemente utilizados na abordagem filtro para a seleção de atributos, descritos a seguir, além do algoritmo FDimBF proposto neste trabalho. É descrito também o algoritmo $\mathcal{C}4.5$ (Quinlan, 1993), o qual é um algoritmo para indução de árvores de decisão. Esse algoritmo realiza seleção embutida de atributos ao construir árvores e regras de decisão.

ReliefF: O algoritmo Relief (Kira and Rendell, 1992) trabalha por meio da amostragem aleatória de exemplos do conjunto de dados e localização do vizinho mais próximo da mesma classe e do vizinho mais próximo da classe oposta. Os valores dos atributos dos vizinhos mais próximos são comparados aos da classe amostrada e utilizados para atualizar os pesos de relevância de cada atributo em relação à classe. Esse

processo é repetido um número m de vezes. A idéia do Relief é que atributos importantes devem diferenciar exemplos de classes diferentes e possuir valores similares para exemplos da mesma classe. A proposta original do algoritmo Relief, a qual permitia trabalhar com duas classes, foi mais tarde estendida no algoritmo ReliefF para lidar com ruído e conjuntos de dados contendo múltiplas classes (Kononenko, 1994). No ReliefF, a influência de ruído nos dados é amenizada por meio da distribuição da contribuição dos k vizinhos mais próximos da mesma classe do exemplo correntemente considerado e de k vizinhos mais próximos de cada uma das classes diferentes do exemplo amostrado, ao invés de considerar apenas um único vizinho mais próximo.

É interessante notar que quanto maior o valor de m, *i.e.* o número de exemplos amostrados a partir do conjunto de dados, mais confiáveis são as estimativas fornecidas pelo algoritmo ReliefF, embora aumentar m signifique aumentar o tempo necessário para a execução desse algoritmo. ReliefF apresenta uma complexidade de tempo de $O(m \cdot N \cdot M)$, onde N é a quantidade de exemplos do conjunto de dados, M e ó número de atributos desse conjunto de dados e m, como mencionado anteriormente, o número de vezes que o algoritmo procura por exemplos no conjunto de dados para calcular os pesos para os atributos (Robnik-Sikonja and Kononenko, 2003).

FCBF: O algoritmo FCBF (*Fast Correlation-Based Filter*) (Yu and Liu, 2004) realiza a seleção de atributos em duas etapas: primeiramente, os atributos são analisados para determinar o subconjunto de atributos relevantes em relação à classe, removendo os atributos irrelevantes; na segunda etapa, por meio da análise de redundância, determina e remove os atributos redundantes a partir do subconjunto que contém apenas os atributos relevantes, produzindo o subconjunto final de atributos selecionados. Esse algoritmo utiliza a medida *Symmetrical Uncertainty* — SU (Press et al., 1992) como a medida de correlação para aproximar tanto a análise de relevância quanto a análise de redundância.

Assim, na primeira etapa, a medida SU entre cada atributo e a classe é calculada para todos os atributos, os quais são classificados de acordo com sua relevância em relação à classe. Apenas os atributos que possuírem um valor SU maior que um limiar mínimo, que determina quão relevantes os atributos devem ser para serem considerados, são analisados na próxima etapa. Na segunda etapa, os atributos são avaliados na ordem em que foram classificados na etapa anterior, de acordo com a redundância de uns em relação aos outros, produzindo um subconjunto final contendo apenas os atributos relevantes e não redundantes. É importante notar que no algoritmo FCBF os atributos numéricos são discretizados utilizando o algoritmo para discretização de atributos *Minimum Description Lenght* — MDL — proposto por Fayyad and Irani (1993).

O FCBF apresenta a vantagem, sobre as abordagens tradicionais para avaliação de subconjuntos de atributos, de que por meio da separação das tarefas de análise de relevância e a redundância, ele evita o alto custo da busca por subconjuntos. Esse algoritmo apresenta uma complexidade de tempo de $O(M^2)$ (Yu and Liu, 2004).

CBF: O algoritmo CBF (*Consistency-Based Filter* — CBF)(Liu and Setiono, 1996) é um algoritmo probabilístico que avalia os subconjuntos de atributos de acordo com sua consistência em relação à classe. Algoritmos que consideram essa medida procuram por combinações de atributos cujos valores particionem os dados em subconjuntos com alguma classe majoritária. Usualmente, a busca favorece subconjuntos pequenos de atributos que apresentam alta consistência com o atributo classe. Na proposta original desse algoritmo, apenas atributos discretos podem ser considerados para análise e os subconjuntos de atributos são determinados utilizando o algoritmo de Las Vegas (Brassard and Bratley, 1997), o qual realiza escolhas probabilísticas para auxiliá-lo na procura mais rápida por soluções corretas. A idéia desse algoritmo é, por um número máximo de tentativas max_tries, gerar subconjuntos de atributos e avaliá-los quanto ao seu tamanho e a sua inconsistência em relação à classe. Ao final, o subconjunto de atributos selecionado será aquele que, dentro do número máximo de tentativas, possuir o menor tamanho e a menor inconsistência. Esse algoritmo apresenta uma complexidade de tempo de $O(max\_tries \cdot N)$ (Liu and Setiono, 1996). Porém, se a busca por subconjuntos de atributos for realizada utilizando *forward selection* ou *backward selection*, esse algoritmo apresenta uma complexidade de tempo de $O(N \cdot M^2)$. Na implementação do algoritmo CBF utilizada neste trabalho, os atributos numéricos também são discretizados utilizando o algoritmo para discretização de atributos *Minimum Description Lenght* — MDL — proposto por Fayyad and Irani (1993).

CFS: O algoritmo CFS (*Correlation-based Feature Selection*) (Hall, 2000) classifica os subconjuntos de atributos de acordo com medidas de avaliação de correlação. Uma das medidas empregadas é a *Symmetrical Uncertainty* (Press et al., 1992). Esse algoritmo é composto, basicamente, por duas etapas: (1) avaliação da correlação entre os atributos e da correlação entre atributos e classe e (2) busca por subconjuntos de atributos e avaliação desses subconjuntos. Desse modo, o CFS considera a habilidade preditiva individual de cada atributo e o grau de correlação entre esses atributos, incluindo a classe. Para a seleção de atributos utilizando o algoritmo CFS e a medida SU, é necessário que os atributos sejam discretizados. Na implementação utilizada neste trabalho é considerado também o algoritmo para discretização de atributos *Minimum Description Lenght* — MDL — proposto por Fayyad and Irani (1993).

Como esse algoritmo avalia subconjuntos de atributos, é necessário definir como a busca será conduzida e o critério de parada para a busca por esses subconjuntos. A

implementação desse algoritmo permite realizar a busca de três maneiras: *forward selection*, *backward selection* e *best first*. Quanto ao critério de parada da busca, nessa implementação foi estipulado que a busca termina após a geração de cinco subconjuntos de atributos que não mostrem melhores resultados quando comparados ao melhor subconjunto corrente. O algoritmo CFS apresenta uma complexidade de tempo de $O(N \cdot M^2)$ (Hall, 1999).

$\mathcal{C}$4.5: O algoritmo $\mathcal{C}$4.5 é um dos sucessores do algoritmo ID3 (Quinlan, 1986), o qual pertence a uma classe mais genérica de algoritmos de aprendizado de máquina denominado *Top Down Induction of Decision Trees* — TDIDT. Um nó em uma árvore de decisão representa um teste de um atributo em particular. De um modo simplificado, a construção de uma árvore de decisão procede do seguinte modo: usando o conjunto de treinamento, um atributo é escolhido para particioná-lo de acordo com o valor desse atributo. Sucessivamente, para cada subconjunto outros atributos são selecionados, segundo algum critério, para particioná-los. Esse processo continua enquanto cada subconjunto contém exemplos pertencentes a classes diferentes. Uma vez que um subconjunto uniforme, *i.e.* todos ou quase todos os exemplos naquele subconjunto pertencem à mesma classe, é obtido, um nó folha é criado e rotulado com o nome da respectiva classe. Diversas extensões foram adicionadas ao ID3, tais como tratamento de atributos numéricos, valores faltantes ou desconhecidos e o uso do critério de razão de ganho ao invés do critério de ganho, usado na versão original do ID3 para selecionar os atributos que particionam os subconjuntos de exemplos. O propósito original do $\mathcal{C}$4.5 não é a seleção de atributos, porém, como o algoritmo realiza seleção embutida de atributos ao construir a árvore de decisão, os atributos utilizados como nós de decisão podem ser interpretados como sendo relevantes em relação à classe e ordenados de acordo com o número de vezes que aparecem nas regras geradas a partir da árvore. O $\mathcal{C}$4.5 apresenta complexidade de tempo de $O(N \cdot logN \cdot M)$ (Witten and Frank, 2000).

FDimBF: O algoritmo FDimBF (*Fractal Dimension-Based Filter*) (Lee et al., 2005,a,b; Lee and Monard, 2003), realiza, assim como o algoritmo FCBF, a seleção de atributos em duas etapas. Na primeira etapa, os atributos relevantes em relação à classe são selecionados. Na segunda etapa, somente os atributos não redundantes são selecionados a partir do subconjunto de atributos escolhidos na etapa anterior. O subconjunto final de atributos selecionados será composto por atributos relevantes em relação à classe e não redundantes entre si — Seção 4 na página 5. Para a análise de relevância, neste trabalho são propostas duas versões desse algoritmo. FDimBF(1) considera uma medida de informação para a seleção de atributos relevantes em relação à classe. Essa medida é implementada por meio da utilização do algoritmo $\mathcal{C}$4.5. Os atributos que pertencem às regras, construídas a partir da árvore de decisão, são considerados como relevantes e classificados em ordem de relevância de acordo com o número de vezes que aparecem nessas regras. Já FDimBF(2)

considera uma medida de distância para selecionar atributos relevantes em relação à classe, a qual é implementada por meio da utilização do algoritmo ReliefF, durante a primeira etapa. Ambos FDimBF(1) e FDimBF(2) consideram a medida de dimensão fractal para a seleção de atributos não redundantes na segunda etapa do processo de seleção de atributos — Seção 3 na página 4. O algoritmo FDimBF apresenta uma complexidade de tempo de $O(N \cdot M^2)$.

É importante notar que as duas versões do algoritmo proposto neste trabalho tratam atributos numéricos sem a necessidade que eles sejam discretizados.

Apesar de FDimBF realizar busca por subconjuntos durante a segunda etapa na seleção de atributos, há dois fatores que contribuem para que sua complexidade em tempo seja comparável a diversos algoritmos encontrados na literatura:

1. durante a primeira etapa, menos custosa, é analisado um número maior de atributos; desse modo, durante a segunda etapa, mais custosa, há em geral uma diminuição do número de atributos a serem analisados e
2. durante a segunda etapa, é utilizado um algoritmo rápido para o cálculo da DF e seleção de atributos não redundantes.

A Tabela 6.1 resume as principais características desses algoritmos e do algoritmo FDimBF proposto neste trabalho.

	$\mathcal{C}4.5$	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
Avaliação Individual	✓	✓		✓		✓	✓
Avaliação de Subconjuntos			✓		✓	✓	✓
Medida de Ganho de Informação	✓					✓	
Medida de Distância		✓					✓
Medida de Correlação			✓	✓		✓	✓
Medida de Consistência					✓		

Tabela 6.1: Características dos algoritmos de SA

7 Configuração dos Experimentos

Os experimentos realizados foram organizados em quatro etapas, as quais são ilustradas na Figura 7.1.

Etapa 1: nessa etapa foram realizadas a limpeza e a preparação dos dados. A tarefa de limpeza dos dados consistiu na remoção de valores desconhecidos da seguinte

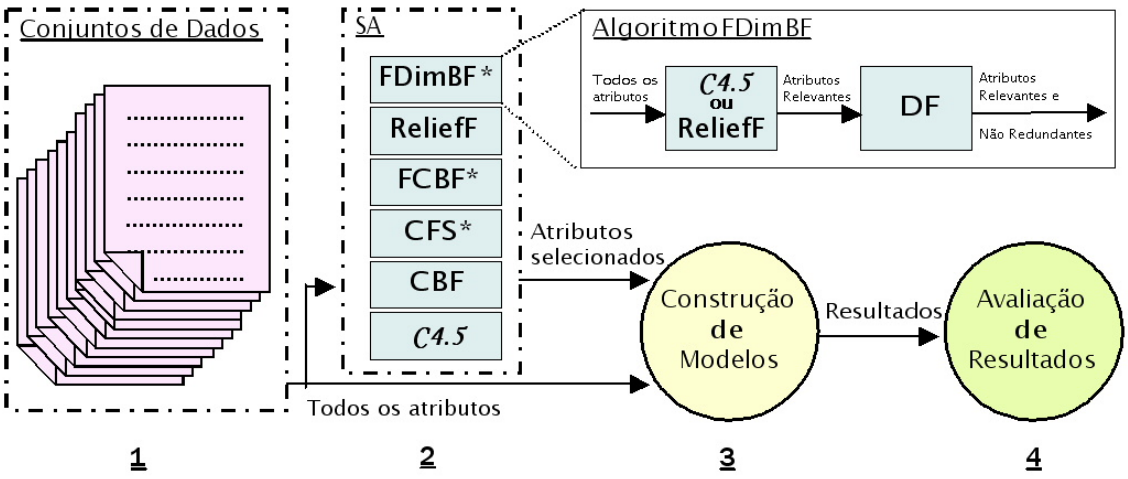


Figura 7.1: Configuração dos experimentos

maneira: para valores desconhecidos concentrados em alguns poucos exemplos, esses exemplos foram removidos, enquanto que para valores desconhecidos concentrados em um atributo, a coluna correspondente foi removida do conjunto de dados. A principal razão para a remoção de valores desconhecidos do conjunto de dados é que alguns dos algoritmos utilizados nesses experimentos tratam valores faltantes de modo especial (Batista and Monard, 2003a), enquanto outros algoritmos não tratam esse tipo de informação. Assim, com o intuito de não introduzir interferências associadas ao uso de um ou outro método para tratar esse problema, foi decidida a remoção de valores desconhecidos do conjunto de dados. Ao final dessa etapa, os dados foram transformados para a sintaxe requerida por cada um dos algoritmos e ferramentas utilizados neste trabalho.

Etapa 2: nessa etapa foi realizada a seleção de atributos utilizando os algoritmos descritos na Seção 6 e o algoritmo por nós proposto — Seção 4. Todos esses algoritmos, a exceção do algoritmo proposto neste trabalho, estão implementados na ferramenta Weka (Witten and Frank, 2000) e foram executados considerando seus parâmetros configurados com os valores padrão. Deve ser observado que os algoritmos marcados com $*$ na Figura 7.1 são aqueles que tratam tanto o problema da relevância de atributos, em relação ao atributo classe, quanto o problema da redundância de atributos.

Etapa 3: nessa etapa foram induzidos os modelos (classificadores) usando todos os atributos remanescentes da Etapa 1 e apenas os atributos selecionados na etapa anterior. Esses modelos foram construídos utilizando o algoritmo $\mathcal{C}4.5$ (Quinlan, 1993).

Etapa 4: nessa última etapa, os resultados foram avaliados por meio da estimativa da média do erro de cada um dos modelos construídos usando validação cruzada com 10 partições (*10 fold cross-validation*). Esse modo de avaliação foi escolhido pois, para conjuntos de dados naturais ou reais, o conhecimento prévio sobre que atributos

são importantes, em geral, não está disponível. Desse modo, a precisão preditiva é comumente utilizada como uma medida indireta para avaliar a qualidade dos atributos selecionados.

Dos 11 conjuntos de dados considerados neste trabalho, somente dois foram submetidos à limpeza de dados: Breast Cancer e Hungarian. O primeiro conjunto de dados possuía originalmente 699 exemplos e 9 atributos. Nesse conjunto de dados os valores faltantes estavam concentrados em alguns poucos exemplos, assim, após a realização dessa tarefa, passou a ser representado por 683 exemplos e o mesmo número de atributos. Já o conjunto de dados Hungarian, o qual continha 294 exemplos descritos por 13 atributos, possuía valores faltantes concentrados tanto em exemplos quanto em atributos. Desse modo, após a limpeza de dados, o novo conjunto de dados Hungarian passou a ser descrito por 261 exemplos e 10 atributos.

Para auxiliar na tarefa de construção dos modelos e avaliação desses modelos por meio de validação cruzada com 10 partições, foi utilizado o ambiente para gerenciamento de experimentos SNIFFER, que faz parte do projeto DISCOVER (Batista, 2003; Batista and Monard, 2003b, 2005, 2002).

O projeto DISCOVER oferece vantagens em relação a outros sistemas com objetivos semelhantes, pois permite a visão unificada que os formatos baseados em padrões proporcionam ao pesquisador (desenvolvedor) de novos componentes. Os padrões de representação foram definidos por área, tendo sido proposta por Prati et al. (2001a) uma sintaxe padrão para representação de conhecimento de diversos indutores simbólicos denominada $\mathcal{PBM}$ (Prati et al., 2002, 2001b). Para a representação de dados foi proposta uma sintaxe padrão (Batista, 2001), denominada *Discover Dataset Sintax* — DSX —, a qual permite a utilização da biblioteca de classes *Discover Object Library* — DOL — (Batista and Monard, 2005), para, entre outras funcionalidades, converter os arquivos de dados para sintaxe utilizada por diversos sistemas de aprendizado simbólico, tais como $\mathcal{C}4.5$, $\mathcal{C}4.5$rules, $\mathcal{CN}2$ entre outros.

8 Resultados e Discussão

Para cada conjunto de dados, foi realizada a seleção de atributos usando as duas versões do algoritmo proposto neste trabalho, *i.e.* FDimBF(1) e FDimBF(2), e os algoritmos $\mathcal{C}4.5$, ReliefF, CFS, CBF e FCBF, totalizando 77 experimentos. Como mencionado anteriormente, foram gerados modelos considerando os atributos selecionados pelos algoritmos citados e também considerando os conjuntos de dados descritos pelos conjuntos originais de atributos (sem SA), totalizando 88 modelos construídos. Os erros dos classificares foram estimados por meio de validação cruzada com 10 partições e foram comparados usando o teste estatístico não paramétrico Kruskal-Wallis para grupos não pareados, seguido do

pós-teste de Dunn[1] (Motulsky, 1995).

Os resultados obtidos, apresentados a seguir, estão organizados da seguinte maneira:

1. Dimensão Fractal e comportamento dos conjuntos de dados quanto à característica fractal;
2. Subconjuntos de atributos selecionados pelos algoritmos considerados neste trabalho, bem como a redução do número de atributos;
3. Distribuição aproximada para cada um dos atributos de cada conjunto de dados;
4. Performance dos algoritmos em relação à precisão e quantidade de atributos selecionados;
5. Análise da significância estatística dos resultados e
6. Características dos conjuntos de dados associadas à utilização da DF como uma medida adequada.

8.1 Dimensão Fractal e Comportamento Fractal dos Conjuntos de Dados

Como mencionado anteriormente, a medida de Dimensão Fractal é utilizada para a determinação de quantos atributos são não redundantes a partir do subconjunto de atributos relevantes de um conjunto de dados. De uma maneira simplificada, a classificação de quais atributos são importantes, segundo a DF, é realizada por meio da determinação dos atributos que, quando retirados do conjunto de dados, causam uma mudança significativa no valor da DF recalculada. Desse modo, os atributos são classificados de acordo com sua importância para o cálculo da Dimensão Fractal.

Para a análise de resultados associados à DF, é importante observar os seguintes três aspectos:

1. o formato da curva de comportamento do conjunto de dados;
2. o número de pontos utilizados para construir essa curva e
3. o ajuste da reta, que determina a DF, em relação a curva.

Dois exemplos, construídos com o auxílio da ferramenta MDE (Seção 4 na página 5), são ilustrados nas Figuras 8.1 e 8.2, correspondentes aos conjuntos de dados Hungarian e Waveform, respectivamente. Nessas figuras é possível observar: (a) curva de comportamento do conjunto de dados, que representa o gráfico em escala logarítmica da soma dos pontos existentes em uma célula de lado r pelo tamanho da célula r, e (b) reta que aproxima o cálculo da Dimensão Fractal desse conjunto de dados.

[1] Testes estatísticos realizados utilizando GraphPad InStat versão 3.06 para Windows, GraphPad Software, http://www.graphpad.com.

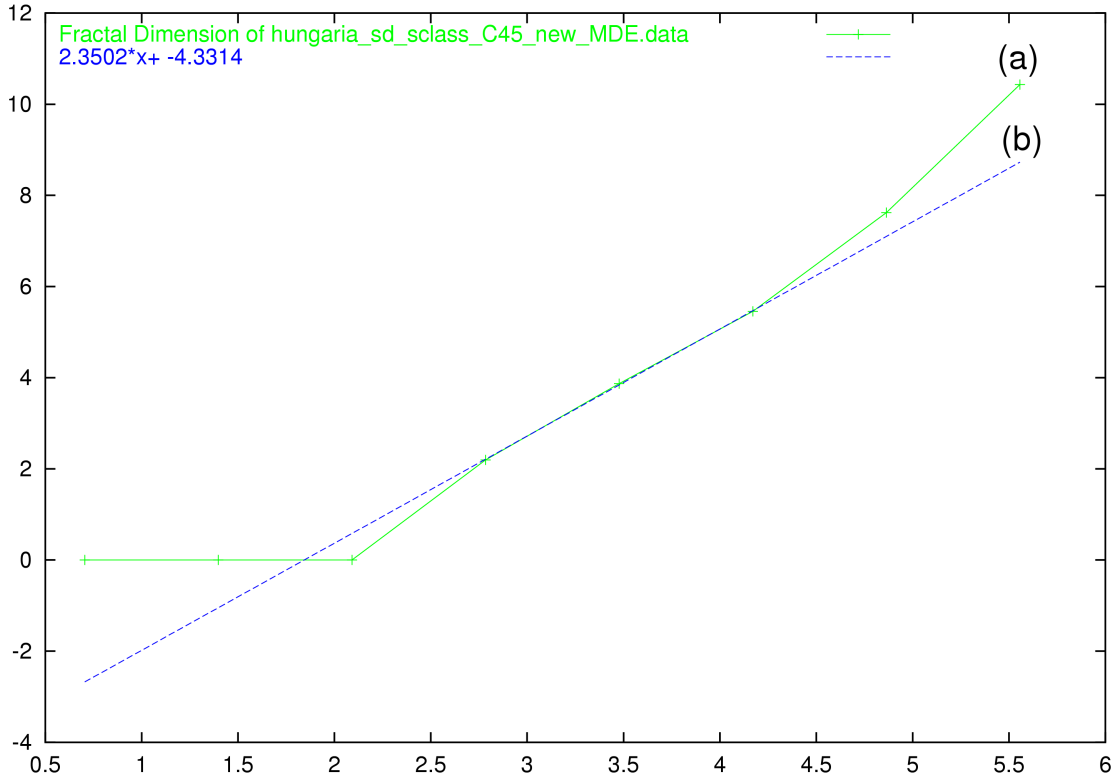


Figura 8.1: Gráfico gerado utilizando o método *Box Count Plot* — Hungarian

A curva (a) da Figura 8.1 para o conjunto de dados Hungarian ilustra o comportamento de um conjunto de dados que apresenta característica de um fractal estatisticamente auto-similar, pois é possível identificar um trecho "bastante" linear na curva de comportamento desse conjunto de dados. Ao fazer essa identificação, é importante também observar o número de pontos utilizados para construir essa curva, sendo que o número mínimo considerado razoável é de três pontos. Em relação ao ajuste da reta ao trecho, aproximadamente linear da curva, é importante também observar se esse ajuste foi realizado sobre um número aceitável de pontos. No exemplo do conjunto de dados Hungarian, a curva foi construída tomando em consideração sete pontos e a reta foi ajustada sobre cinco pontos. Por outro lado, a curva (a) da Figura 8.2 para o conjunto de dados Waveform representa um caso em que o conjunto de dados parece não exibir característica fractal, pois apresenta um trecho em que praticamente não há variação na quantidade de pontos dentro de células de diferentes tamanhos, apresentando somente no trecho final da curva uma pequena variação. Esse comportamento do conjunto de dados também é refletido no ajuste da reta da DF o qual, para o conjunto de dados Waveform, foi realizado sobre apenas dois pontos.

Os gráficos de comportamento relacionado à DF dos 11 conjuntos de dados são apresentados no Apêndice A na página 41 deste trabalho.

A Tabela 8.1 apresenta as informações associadas ao cálculo da DF para cada um dos 11 conjuntos de dados utilizados nos experimentos, onde:

- Algoritmo: indica a abordagem do algoritmo proposto, *i.e.*, se a seleção de atributos

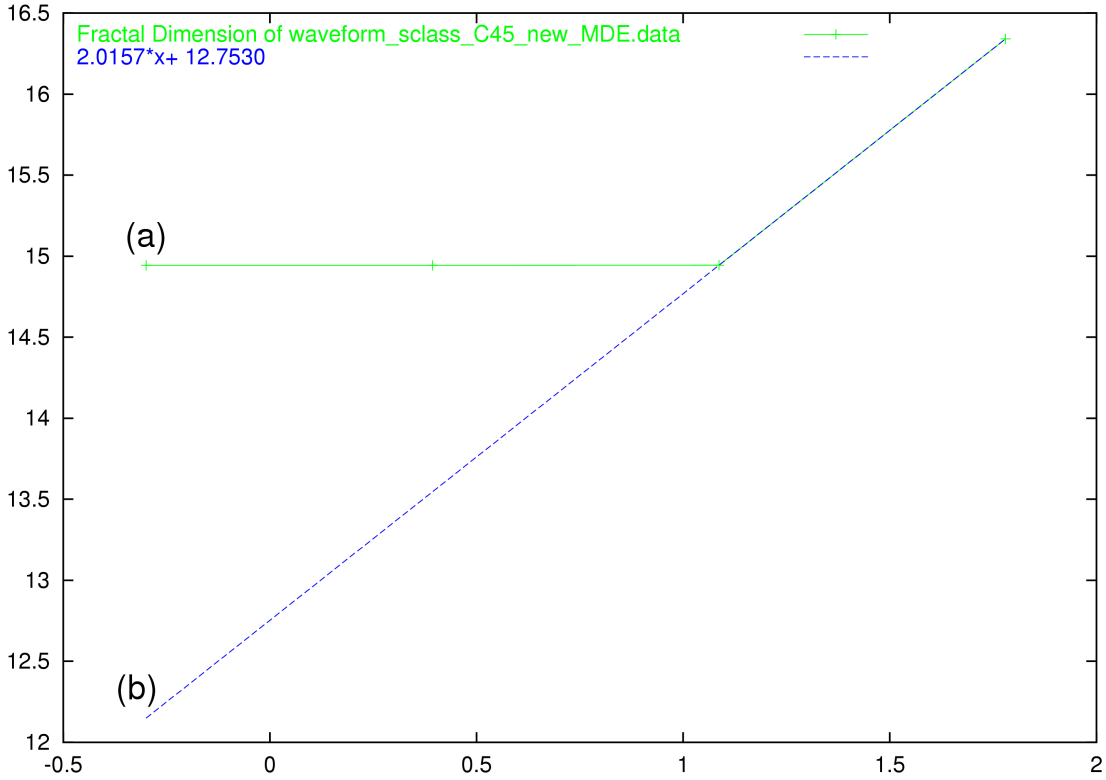


Figura 8.2: Gráfico gerado utilizando o método *Box Count Plot* — Waveform

relevantes em relação à classe foi realizada previamente aplicando a medida de ganho de informação (FDimBF(1)) ou a medida de distância (FDimBF(2));

- # Atrib. Orig.: número de atributos após a remoção de valores desconhecidos do conjunto de dados[2];
- # Ex.: número de exemplos após a remoção de valores desconhecidos do conjunto de dados[3];
- # Atrib. Relev.: número de atributos relevantes selecionados a partir da aplicação das medidas de ganho de informação (FDimBF(1)) ou de distância (FDimBF(2));
- DF: dimensão fractal do conjunto de dados, considerando somente os atributos relevantes em relação à classe;
- # Atributos Selecionados: número de atributos selecionados utilizando a DF como medida de redundância;
- # Pontos (Curva): número de pontos utilizados para a construção da curva de comportamento do conjunto de dados e
- # Pontos (Reta): número de pontos utilizados para ajustar a reta sobre a curva de comportamento do conjunto de dados.

[2] Apenas o conjunto de dados Hungarian apresentou valores faltantes concentrados em atributos.

[3] Ambos os conjuntos de dados Breast Cancer e Hungarian apresentaram valores faltantes concentrados em exemplos.

Conjunto de Dados	Algoritmo	# Atrib. Orig.	# Ex.	# Atrib. Relev.	DF	# Atrib. Selec.	# Pontos (Curva)	# Pontos (Reta)
Breast Cancer		9	683					
	FDimBF(1)			7	2,20	3	4	3
	FDimBF(2)			9	2,30	3	4	3
Bupa		6	345					
	FDimBF(1)			6	3,79	4	6	4
	FDimBF(2)			5	3,42	4	7	4
German		24	1000					
	FDimBF(1)			24	11,46	12	6	2
	FDimBF(2)			24	11,39	12	6	2
Hungarian		10	261					
	FDimBF(1)			9	2,35	3	7	5
	FDimBF(2)			10	3,60	4	5	4
Ionosphere		34	351					
	FDimBF(1)			15	2,79	3	5	5
	FDimBF(2)			33	3,23	4	4	3
Pima		8	769					
	FDimBF(1)			7	2,75	3	5	3
	FDimBF(2)			8	3,14	4	5	3
Satimage		36	4435					
	FDimBF(1)			36	5,09	6	4	3
	FDimBF(2)			36	5,09	6	4	3
Segment		19	2310					
	FDimBF(1)			16	3,07	4	8	4
	FDimBF(2)			18	3,07	4	8	4
Sonar		60	208					
	FDimBF(1)			15	4,95	5	3	3
	FDimBF(2)			60	9,54	10	2	2
Vehicle		18	846					
	FDimBF(1)			18	5,83	6	6	4
	FDimBF(2)			18	5,83	6	6	4
Waveform		21	5000					
	FDimBF(1)			21	2,02	3	3	2
	FDimBF(2)			21	2,02	3	3	2

Tabela 8.1: Informações associadas à dimensão fractal dos conjuntos de dados

Dos 11 conjuntos de dados considerados neste trabalho, apenas para Sonar, quando utilizado com FDimBF(2), *i.e.* ReliefF, não foi possível construir uma curva de comportamento do conjunto de dados com um mínimo de três pontos. Quanto ao número de pontos usados para o ajuste da reta para o cálculo da DF, em dois casos, conjuntos de dados German e Waveform, a reta foi ajustada com menos de três pontos para FDimBF(1) e em três casos, conjuntos de dados German, Sonar e Waveform, para FDimBF(2).

O resultado da análise dos gráficos de comportamento dos conjuntos de dados quanto a sua característica fractal é apresentado na Tabela 8.2, onde:

- # Pontos (Curva, Reta): mostra, respectivamente, o número de pontos utilizado pelo MDE para interpolar a curva e a reta de ajuste correspondente e

- Fractal: classifica a característica fractal do conjunto de dados como Muito Bom, Bom, Mediano e Ruim considerando o formato da curva de comportamento do conjunto de dados e o número de pontos usado para construir essa curva.

Algoritmo	Conjunto de Dados	# Pontos (Curva, Reta)	Caract. Fractal	Conjunto de Dados	# Pontos (Curva, Reta)	Caract. Fractal
	Breast			Satimage		
FDimBF(1)	Cancer	(4, 3)	Bom		(4, 2)	Mediano
FDimBF(2)		(4, 3)	Bom		(4, 2)	Mediano
	Bupa			Segment		
FDimBF(1)		(6, 4)	Muito Bom		(8, 4)	Bom
FDimBF(2)		(7, 4)	Muito Bom		(8, 4)	Bom
	German			Sonar		
FDimBF(1)		(6, 2)	Ruim		(3, 3)	Muito Bom
FDimBF(2)		(6, 2)	Ruim		(2, 2)	Ruim
	Hungarian			Vehicle		
FDimBF(1)		(7, 5)	Muito Bom		(6, 4)	Bom
FDimBF(2)		(5, 4)	Muito Bom		(6, 4)	Bom
	Ionosphere			Waveform		
FDimBF(1)		(5, 5)	Muito Bom		(3, 2)	Ruim
FDimBF(2)		(4, 3)	Muito Bom		(3, 2)	Ruim
	Pima					
FDimBF(1)		(5, 3)	Bom			
FDimBF(2)		(5, 3)	Bom			

Tabela 8.2: Resultado da análise dos gráficos de comportamento dos conjuntos de dados quanto à característica fractal

Uma análise dos gráficos de comportamento dos conjuntos de dados mostrou que em relação à característica fractal houve quatro e três Muito Bom, cinco e cinco Bom, um e um Mediano e dois e três Ruim para FDimBF(1)e FDimBF(2), respectivamente.

8.2 Subconjuntos de Atributos Selecionados

Dois principais fatores, além das características próprias do conjunto de dados, podem influenciar no subconjunto de atributos selecionado por algoritmos de seleção de atributos e estão relacionados com:

- a avaliação dos atributos, *i.e.* se os atributos são avaliados individualmente ou considerando um subconjunto de atributos e
- a medida utilizada para determinar a importância dos atributos.

Como mencionado anteriormente, neste trabalho foram considerados quatro algoritmos freqüentemente citados na literatura para a seleção de atributos e o algoritmo $\mathcal{C}4.5$, além do algoritmo FDimBF proposto —- Tabela 6.1 na página 13. Três desses algoritmos ($\mathcal{C}4.5$, ReliefF e FCBF) realizam a seleção utilizando o critério de avaliação individual de atributos e os outros dois (CFS e CBF) o critério de avaliação de subconjuntos de atributos. O algoritmo FDimBF realiza a primeira parte da seleção de atributos por meio de avaliação individual de atributos e a segunda parte por meio de avaliação de subconjuntos de atributos. Em relação à medida utilizada para determinar a importância dos atributos, esses algoritmos usam medidas de distância (ReliefF e FDimBF(2)), correlação (CFS, FCBF e FDimBF(1)), ganho de informação ($\mathcal{C}4.5$) e consistência (CFS).

A Tabela 8.3 apresenta um resumo da quantidade de atributos selecionados por cada um dos algoritmos e suas respectivas percentagens. Também é apresentada essa informação para o algoritmo $\mathcal{C}4.5$, o qual é utilizado na etapa de seleção de atributos relevantes em relação à classe no algoritmo FDimBF(1). As informações dessa tabela estão organizadas do seguinte modo:

Na primeira coluna é apresentado o conjunto de dados ao qual referem-se as informações. Na segunda coluna é indicada a quantidade original, *i.e.* depois da remoção de valores desconhecidos, de atributos de cada conjunto de dados. Para cada um deles, na primeira linha são descritos o número de atributos referentes ao subconjunto selecionado por cada um dos algoritmos e na segunda linha é apresentada a respectiva percentagem. As últimas duas linhas mostram a média de atributos selecionados por cada algoritmo (Média # Atrib.) e a respectiva percentagem (Média % Atrib.).

	Sem SA	$\mathcal{C}4.5$	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
Breast Cancer	9	7	9	9	9	7	3	3
		77,78	100,00	100,00	100,00	77,78	33,33	33,33
Bupa	6	6	5	1	1	1	4	4
		100,00	83,33	16,67	16,67	16,67	66,67	83,33
German	24	24	24	2	15	15	12	12
		100,00	100,00	8,33	62,50	62,50	50,00	50,00
Hungarian	10	9	10	3	6	5	3	4
		90,00	100,00	30,00	60,00	50,00	30,00	40,00
Ionosphere	34	15	33	14	33	7	3	4
		44,12	97,06	41,18	97,06	20,59	8,82	11,76
Pima	8	7	8	3	8	8	3	4
		87,50	100,00	37,50	100,00	100,00	37,50	50,00
Satimage	36	36	36	23	36	12	6	6
		100,00	100,00	63,89	100,00	33,33	16,67	16,67
Segment	19	16	18	5	18	9	4	4
		84,21	94,74	26,32	94,74	47,37	21,05	21,05
Sonar	60	15	60	19	21	14	5	10
		25,00	100,00	31,67	35,00	23,33	8,33	16,67
Vehicle	18	18	18	11	18	18	6	6
		100,00	100,00	61,11	100,00	100,00	33,33	33,33
Waveform	21	21	21	15	19	12	3	3
		100,00	100,00	71,43	90,48	57,14	14,29	14,29
Média # Atrib.		16	22	10	17	10	5	5
Média % Atrib.		82,60	97,74	44,37	77,86	53,52	29,09	32,16

Tabela 8.3: Resumo da quantidade de atributos selecionados por cada um dos algoritmos e suas respectivas percentagens

Considerando somente o tamanho dos subconjuntos de atributos selecionados por cada algoritmo, o algoritmo ReliefF foi o que selecionou os maiores subconjuntos de atributos, variando de um mínimo de $83,33\%$ do total de atributos para o conjunto de dados Bupa até o máximo de $100,00\%$ (todos os atributos) para oito do total de 11 conjuntos de dados. O algoritmo CFS selecionou o menor número de atributos, em relação ao número de atributos selecionados pelos outros algoritmos, em um conjunto de dados: German $(8,33\%)$. Para outros três conjuntos de dados, o algoritmo CFS selecionou juntamente com outros algoritmos o menor número de atributos: Bupa $(16,67\%)$ em conjunto com

FCBF e CBF e Hungarian (30,00%) e Pima (37,50%) em conjunto com FDimBF(1). O algoritmo FDimBF, nas duas versões, foi o que mais freqüentemente selecionou os menores subconjuntos de atributos: cinco vezes FDimBF(1) em conjunto com FDimBF(2), duas vezes FDimBF(1) sozinho para os conjuntos de dados Ionosphere (8,82%) e Sonar (8,33%) e duas vezes FDimBF(1) em conjunto com o algoritmo CFS, como mencionado anteriormente.

As Figuras 8.3a a 8.3k mostram graficamente, para cada conjunto de dados, o número de atributos selecionados e a respectiva percentagem *versus* o algoritmo de SA. Entre parênteses é apresentada a média da percentagem de atributos selecionados pelos algoritmos de SA para cada conjunto de dados.

É interessante observar que dentre os 11 conjuntos de dados, em cinco deles, Breast Cancer, Satimage, Segment, Vehicle e Waveform, as abordagens FDimBF(1) e FDimBF(2) selecionaram o mesmo subconjunto final de atributos. Para dois conjuntos de dados, Ionosphere e Sonar, as duas abordagens do algoritmo FDimBF selecionaram subconjuntos de atributos totalmente diferentes. No restante dos conjuntos de dados, os subconjuntos de atributos selecionados incluem alguns dos atributos escolhidos por FDimBF(1) e FDimBF(2) em comum. As tabelas que mostram os atributos selecionados por cada um dos algoritmos considerados neste trabalho são apresentadas no Apêndice B.

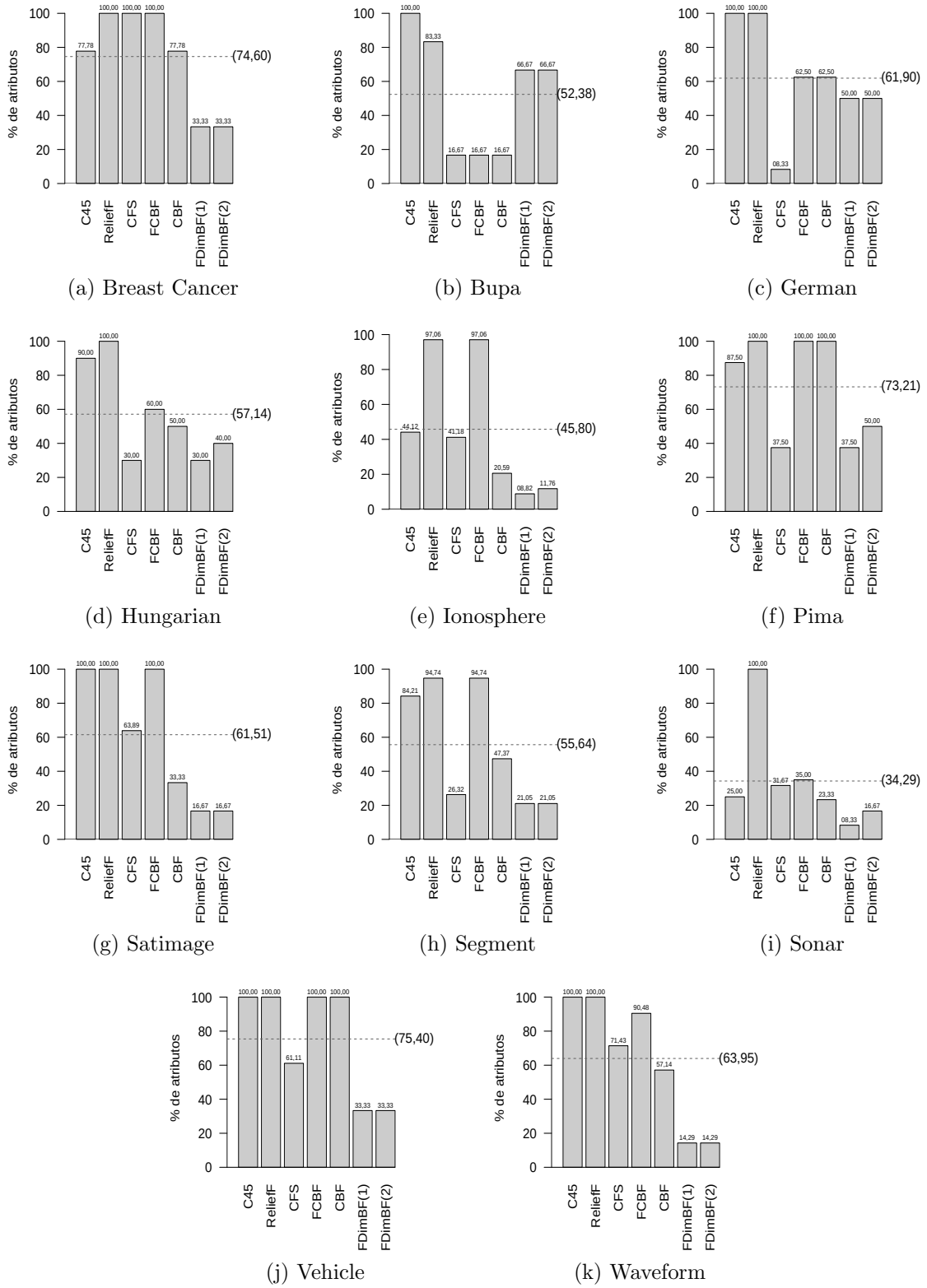
8.3 Formatos Aproximados de Distribuição dos Valores dos Atributos em Relação aos Atributos Selecionados pelo Algoritmo FDimBF

Com o intuito de verificar se a distribuição dos valores dos atributos originais exerce alguma influência sobre os subconjuntos de atributos selecionados por FDimBF(1) e (2), esses três conjuntos de atributos foram analisados do ponto de vista do formato aproximado da distribuição da maioria dos atributos presentes neles e classificados de acordo com seis tipos, T_1, T_2, T_3, T_4, T_5 e T_6, como é mostrado na Figura 8.4. Além disso, uma outra relação de interesse é saber se os algoritmos FDimBF selecionam, preferencialmente, atributos cujos valores obedecem a algum tipo de distribuição.

A Tabela 8.4 apresenta, para cada conjunto de dados e abordagem de FDimBF, em que tipo de formato aproximado de distribuição os atributos podem ser classificados. As distribuições dos valores dos atributos, para cada um dos conjuntos de dados considerados neste trabalho, são apresentadas no Apêndice C.

Dos 11 conjuntos de dados considerados neste trabalho, sete deles possuem a maioria dos atributos com formato aproximado de distribuição do tipo T_3, três do tipo T_1 e um do tipo T_2.

Em relação aos atributos selecionados pelos algoritmos FDimBF, é interessante notar que do total de 11 conjuntos de dados considerados, em 10 deles, a maioria dos atributos apresentaram formatos aproximados de distribuição dos valores semelhantes para



(a) Breast Cancer (b) Bupa (c) German

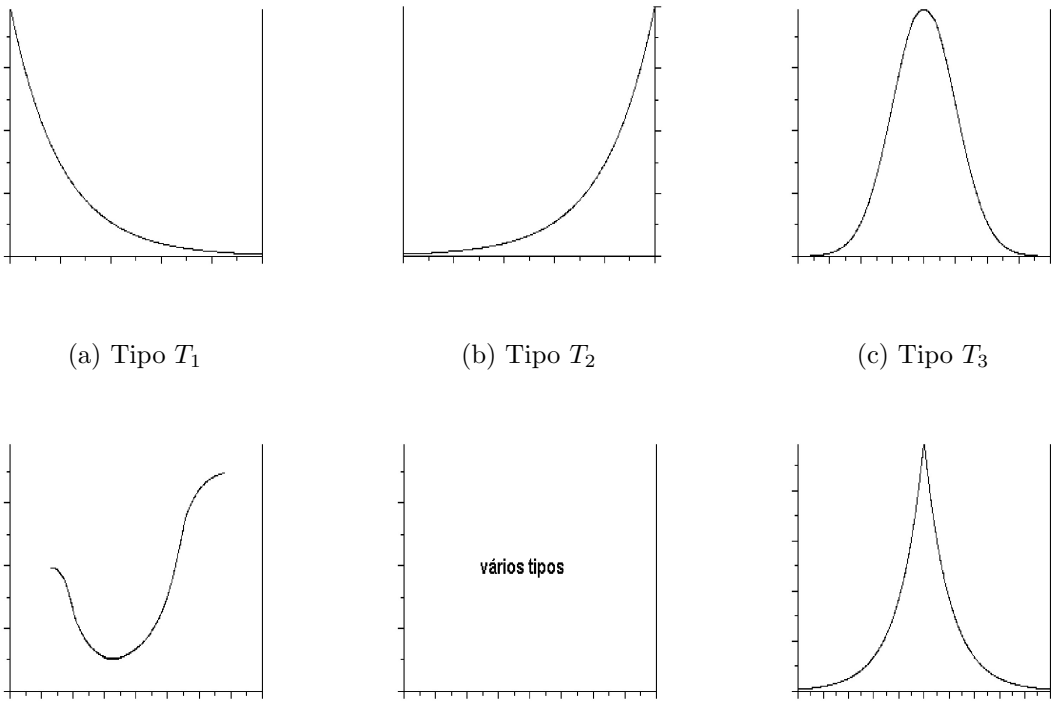
(d) Hungarian (e) Ionosphere (f) Pima

(g) Satimage (h) Segment (i) Sonar

(j) Vehicle (k) Waveform

Figura 8.3: Número de atributos selecionados e a respectiva percentagem *versus* o algoritmo de SA

FDimBF(1) e FDimBF(2). Desses 10 conjuntos de dados, em cinco deles isso se deve aos subconjuntos de atributos selecionados pelas duas abordagens serem os mesmos, como mencionado anteriormente. Nos outros seis conjuntos, quatro deles, Bupa, German, Hungarian e Pima, apresentam intersecção entre os subconjuntos e dois, Ionosphere e Sonar,



(a) Tipo T_1 (b) Tipo T_2 (c) Tipo T_3

(d) Tipo T_4 (e) Tipo T_5 (f) Tipo T_6

Figura 8.4: Tipos de formatos aproximados das distribuições dos valores dos atributos

Conjunto de Dados	Todos os Atributos	Atributos Selecionados por (FDimBF(1), FDimBF(2))
Breast Cancer	T_1	(T_1,T_1)
Bupa	T_3	(T_3,T_3)
German	T_1	(T_4,T_4)
Hungarian	T_3	(T_5,T_3)
Ionosphere	T_2	(T_3,T_3)
Pima	T_3	(T_3,T_3)
Satimage	T_3	(T_3,T_3)
Segment	T_1	(T_5,T_5)
Sonar	T_3	(T_3,T_3)
Vehicle	T_3	(T_5,T_5)
Waveform	T_3	(T_3,T_3)

Tabela 8.4: Formatos da distribuição aproximada dos valores dos atributos

apresentam subconjuntos de atributos totalmente diferentes.

Essa classificação dos subconjuntos de atributos quanto ao formato da distribuição dos valores da maioria de seus atributos mostra que mais de 50% dos subconjuntos apresenta uma distribuição do tipo T_3 de seus valores.

8.4 Performance dos Algoritmos em Relação à Precisão e a Quantidade de Atributos Selecionados

Os resultados dos experimentos foram também avaliados quanto à relação entre a quantidade de atributos selecionados e o erro dos modelos construídos (Tabela 8.5).

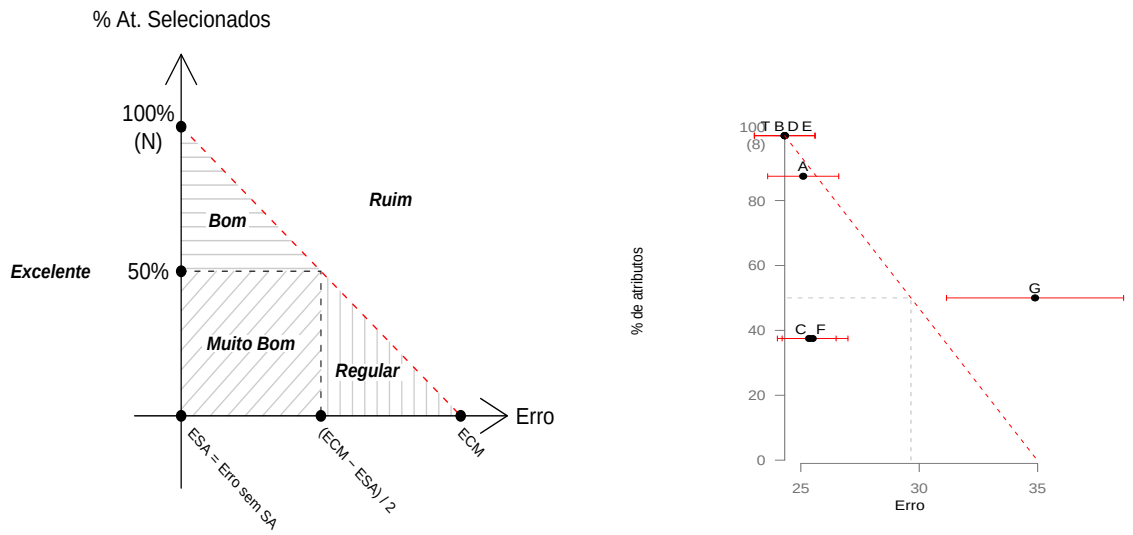
	Sem SA	$\mathcal{C}4.5$	ReliefF	CFS
BreastCancer	5,27 ± 1,03	4,83 ± 0,54	5,72 ± 0,97	4,54 ± 0,70
Bupa	29,57 ± 2,38	32,47 ± 2,58	33,63 ± 3,11	36,77 ± 2,72
German	26,60 ± 1,37	27,00 ± 1,41	25,30 ± 1,07	28,00 ± 0,68
Hungarian	23,40 ± 2,05	21,87 ± 2,01	24,93 ± 1,95	21,48 ± 2,89
Ionosphere	9,97 ± 1,96	11,40 ± 0,85	10,55 ± 2,00	10,27 ± 0,98
Pima	24,32 ± 1,28	25,10 ± 1,50	25,22 ± 1,43	25,35 ± 1,14
Satimage	14,05 ± 0,43	14,57 ± 0,44	13,71 ± 0,62	13,66 ± 0,49
Segment	3,03 ± 0,35	3,46 ± 0,54	3,29 ± 0,30	3,59 ± 0,25
Sonar	24,05 ± 3,70	24,95 ± 2,97	22,12 ± 2,79	23,95 ± 2,64
Vehicle	26,95 ± 1,16	26,72 ± 1,57	25,90 ± 1,62	31,68 ± 1,50
Waveform	23,80 ± 0,51	23,00 ± 0,67	23,70 ± 0,44	22,38 ± 0,41
	FCBF	CBF	FDimBF(1)	FDimBF(2)
BreastCancer	4,54 ± 0,70	4,98 ± 0,62	4,40 ± 0,54	4,40 ± 0,54
Bupa	36,77 ± 2,72	36,77 ± 2,72	42,01 ± 1,37	33,03 ± 2,17
German	27,50 ± 0,81	26,40 ± 1,90	25,50 ± 1,49	26,30 ± 0,79
Hungarian	23,38 ± 2,78	23,40 ± 1,97	24,53 ± 2,45	22,21 ± 2,18
Ionosphere	10,55 ± 2,00	11,40 ± 1,86	19,38 ± 2,48	19,36 ± 1,72
Pima	26,02 ± 1,53	27,82 ± 1,14	25,50 ± 1,49	34,89 ± 3,74
Satimage	14,36 ± 0,65	13,55 ± 0,53	16,80 ± 0,56	16,80 ± 0,56
Segment	3,29 ± 0,30	3,51 ± 0,30	6,15 ± 0,35	6,15 ± 0,35
Sonar	26,38 ± 1,71	25,52 ± 4,27	38,02 ± 2,37	34,55 ± 3,42
Vehicle	27,19 ± 1,56	28,37 ± 1,28	33,92 ± 1,00	33,92 ± 1,00
Waveform	23,24 ± 0,69	24,86 ± 0,88	35,16 ± 0,78	35,16 ± 0,78

Tabela 8.5: Média de erro e erro padrão para cada conjunto de dados e algoritmo considerado

Essa relação é representada graficamente com o objetivo de auxiliar na avaliação da performance dos algoritmos considerando ambas as medidas, como mostrado na Figura 8.5a, sendo:

- Eixo X: representa a percentagem de atributos selecionados em relação ao total de atributos (apresentado entre parênteses) e
- Eixo Y: representa a média do erro, obtido usando validação cruzada com 10 partições.

Nesse gráfico, para cada conjunto de dados, os algoritmos de SA são classificados quanto ao seu posicionamento em relação à percentagem de atributos selecionados e a média do erro e o erro padrão do modelo construído considerando os atributos selecionados por esses algoritmos, dentro de cinco regiões definidas. Primeiramente, duas grandes áreas são delimitadas pela reta que liga o ponto 100% (número total de atributos do conjunto de dados) no eixo X ao ponto ECM no eixo Y, sendo ECM igual ao Erro da Classe Majoritária caso seja menor que 50%, ou igual a 50% caso contrário. Nesse modelo de



(a) (b)

Figura 8.5: Relação entre percentagem de atributos selecionados, média do erro e erro padrão dos modelos construídos: (a) Modelo geral e (b) Conjunto de dados Pima

avaliação, considerou-se que essa reta representa uma proporção mínima entre o que se espera em termos da relação entre percentagem de atributos selecionados e média do erro do modelo construído considerando os atributos selecionados. Assim, qualquer modelo construído com os atributos selecionados por um algoritmo de SA que esteja localizado na região acima dessa reta pode ser considerado de performance Ruim (▼). Abaixo dessa reta e delimitadas pelos eixos X e Y, outras três regiões foram por nós definidas:

- Muito Bom (▲▲): retângulo que delimita a região que corresponde a 50% ou menos de atributos selecionados e até 50% da diferença entre ECM e o erro do modelo construído considerando todos os atributos — ESA —, *i.e.* sem a realização de seleção de atributos;
- Bom (▲): região acima da região Muito Bom e
- Regular (◇): região ao lado direito da região Muito Bom.

Uma quarta região, denominada Excelente (▲▲▲), foi definida como sendo a área à esquerda do eixo X. Assim, qualquer algoritmo que permita a seleção de subconjuntos de atributos que melhorem a precisão do modelo construído é considerado de performance excelente. Nos casos em que o conjunto de atributos selecionados foi igual ao conjunto original de atributos do conjunto de dados, o algoritmo foi classificado como Todos os Atributos Selecionados (—).

Na Figura 8.5b é apresentado um exemplo do modelo de avaliação por nós proposto para o conjunto de dados Pima. Nessa figura é possível identificar a média do erro e o erro padrão considerando o conjunto original de atributos, denominado T, e as posições dos algoritmos, os quais são representados pelas siglas na Tabela 8.6, dentro das cinco regiões descritas anteriormente. Para esse conjunto de dados, o modelo construído utilizando

o subconjunto de atributos selecionado por $\mathcal{C}4.5$ foi considerado bom. Já os algoritmos ReliefF, FCBF e CBF selecionaram todos os atributos do conjunto de dados como sendo importantes. O modelo construído utilizando o subconjunto de atributos selecionado por FDimBF(2) foi considerado ruim, pois encontra-se na região acima da reta definida pelos pontos 100% de atributos selecionados e ECM. Já para a seleção de atributos utilizando os algoritmos CFS e FDimBF(1), os modelos construídos foram considerados muito bons.

Sigla	Algoritmo
A	$\mathcal{C}4.5$
B	ReliefF
C	CFS
D	FCBF
E	CBF
F	FDimBF(1)
G	FDimBF(2)

Tabela 8.6: Algoritmos presentes nos gráficos

A Tabela 8.7 mostra um resumo da classificação dos algoritmos de SA para cada conjunto de dados quanto ao posicionamento dentro das regiões definidas — Figura 8.5a. Para cada conjunto de dados é ainda apresentada, na última coluna — CRes —, uma classificação do resultado da aplicação dos algoritmos de SA indicada por $\uparrow$ (número de classificações Excelente, Muito Bom e Bom maior ou igual a cinco), $\downarrow$ (número de classificações Todos os Atributos Selecionados representa em torno de 50% dos casos) e $\sim$ (maioria das classificações Regular e Ruim). Nas últimas linhas dessa tabela é mostrado um resumo da quantidade de vezes em que o respectivo algoritmo foi classificado como tendo apresentado desempenho Excelente, Muito Bom, Bom, Regular, Ruim e Todos os Atributos Selecionados.

Os algoritmos de SA contribuíram para a redução do número de atributos selecionados em relação ao conjunto original de atributos em seis, identificados por $\uparrow$, dos 11 conjuntos de dados considerados neste trabalho, *i.e.* houve cinco ou mais casos classificados como Excelente, Muito Bom ou Bom. Para quatro conjuntos de dados, identificados por $\sim$, a aplicação dos algoritmos de SA não promoveu a redução dos subconjuntos de atributos selecionados em 50% dos casos, embora para todos eles, os outros 50% dos casos tenham sido classificados como Excelente, Muito Bom ou Bom. Apenas em um caso, identificado por $\downarrow$, cinco modelos construídos utilizando os subconjuntos selecionados pelos algoritmos de SA foram classificados como Regular e Ruim.

Considerando cada algoritmo de SA em relação aos tipos de classificação, os algoritmos CFS e CBF foram os que obtiveram o maior número de classificações excelentes, cada um deles tendo obtido quatro. Quanto às classificações muito boas, FDimBF(1) e FDimBF(2) obtiveram sete e seis, respectivamente. Classificações boas e regulares ocorreram de um modo uniforme entre todos os algoritmos. O algoritmo ReliefF juntamente com as duas versões de FDimBF foram os únicos a apresentar classificações ruins. Ressalta-se que as duas versões do algoritmo FDimBF foram os algoritmos que obtiveram o maior número,

Algoritmo	$\mathcal{C}4.5$	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)	CRes
Breast Cancer	▲▲▲	—	—	—	▲▲▲	▲▲▲	▲▲▲	∼
Bupa	—	▼	◇	◇	◇	▼	▲	↓
German	—	—	▲▲	▲	▲▲▲	▲▲▲	▲▲▲	↑
Hungarian	▲▲▲	—	▲▲▲	▲▲▲	▲▲▲	▲▲	▲▲▲	↑
Ionosphere	▲▲	▲	▲▲	▲	▲▲	▲▲	▲▲	↑
Pima	▲	—	▲▲	—	—	▲▲	▼	∼
Satimage	—	—	▲▲▲	—	▲▲▲	▲▲	▲▲	∼
Segment	▲	▲	▲▲	▲	▲▲	▲▲	▲▲	↑
Sonar	▲▲	—	▲▲▲	▲▲	▲▲	◇	▲▲	↑
Vehicle	—	—	▲	—	—	▲▲	▲▲	∼
Waveform	—	—	▲▲▲	▲▲▲	▲	▲▲	▲▲	↑
Excelente (▲▲▲)	2	0	4	2	4	2	3	
Muito (▲▲)	2	0	4	1	3	7	6	
Bom (▲)	2	2	1	3	1	0	1	
Regular (◇)	0	0	1	1	1	1	0	
Ruim (▼)	0	1	0	0	0	1	1	
Todos os Atributos Selecionados (—)	5	8	1	4	2	0	0	

Tabela 8.7: Classificação dos algoritmos em relação a percentagem de atributos selecionados × erro do modelo construído

nove, de classificações excelente e muito bom, seguidas por CFS e CBF, cada um com oito e sete classificações desses tipos, respectivamente. É interessante observar que o algoritmo ReliefF foi o que apresentou maior número, oito, de seleções de subconjuntos iguais aos conjuntos originais de atributos (não houve redução do número de atributos selecionados) e que os algoritmos FDimBF(1) e FDimBF(2) foram os únicos a promover redução do número de atributos selecionados para todos os conjuntos de dados.

Do total de 77 classificações (11 conjuntos de dados × sete algoritmos de SA), 17 foram excelentes, 23 muito boas, 10 boas, quatro regulares, três ruins e 20 selecionaram todos os atributos do conjunto original de atributos. É possível observar que $64,94\%$ das classificações foram excelentes, muito boas ou boas, $25,97\%$ dos subconjuntos de atributos selecionados foram iguais aos conjuntos originais de atributos e apenas $9,09\%$ foram regulares ou ruins, tendo portanto a maioria dos algoritmos de SA contribuído, utilizando os subconjuntos de atributos selecionados, para a melhoria, quer em relação à redução do número de atributos quer em relação à precisão dos modelos construídos no modelo de classificação proposto — Figura 8.5a.

8.5 Análise da Significância Estatística dos Resultados

Como mencionado anteriormente, para cada conjunto de dados, os algoritmos foram comparados entre si quanto à média do erro medido por meio de validação cruzada com 10 partições utilizando o teste não paramétrico Kruskal-Wallis para grupos não pareados, seguido do pós-teste de Dunn, e nível de significância de 0,05. Esses resultados foram comparados também em relação ao número de vezes que cada algoritmo, para um determinado conjunto de dados, seleciona menos atributos com uma média de erro sem

diferença estatística.

A Tabela 8.9 apresenta, para cada conjunto de dados sem SA (Orig.) e cada um dos subconjuntos selecionados pelos algoritmos considerados neste trabalho — ReliefF, $\mathcal{C}4.5$, CFS, FBCF, CBF e as duas versões de FDimBF representadas por DF(1) e DF(2) nessa tabela — o número original de atributos e o número de atributos selecionados por cada um desses algoritmos. As comparações entre as médias dos erros dos modelos construídos que apresentaram diferença estatisticamente significativa estão em negrito. As siglas utilizadas para representar os conjuntos de dados são apresentadas na Tabela 8.8.

Sigla	Conjunto de Dados
BCa	Breast Cancer
Bup	Bupa
Ger	German
Hun	Hungarian
Ion	Ionosphere
Pim	Pima
Sat	Satimage
Seg	Segment
Son	Sonar
Veh	Vehicle
Wav	Waveform

Tabela 8.8: Siglas para os conjuntos de dados

	BCa	Bup	Ger	Hun	Ion	Pim	Sat	Seg	Son	Veh	Wav
Orig.-$\mathcal{C}4.5$	9–7	6–6	24–24	10–9	34–15	8–7	36–36	19–16	60–15	18–18	21–21
Orig.-ReliefF	9–9	6–5	24–24	10–10	34–33	8–8	36–36	19–18	60–60	18–18	21–21
Orig.-CFS	9–9	6–1	24–2	10–3	34–14	8–3	36–23	19–5	60–19	18–11	21–15
Orig.-FCBF	9–9	6–1	24–15	10–6	34–33	8–8	36–36	19–18	60–21	18–18	21–19
Orig.-CBF	9–7	6–1	24–15	10–5	34–7	8–8	36–12	19–9	60–14	18–18	21–12
Orig.-DF(1)	9–3	**6–4**	24–12	10–3	34–3	8–3	36–6	**19–4**	60–5	18–6	**21–3**
Orig.-DF(2)	9–3	6–4	24–12	10–4	34–4	8–4	**36–6**	**19–4**	60–10	**18–6**	**21–3**

Tabela 8.9: Comparação entre o número original de atributos e o número de atributos selecionados pelos algoritmos de SA. Comparação entre médias de erros dos modelos construídos (em negrito resultados estatisticamente significativos)

O algoritmo ReliefF foi o que reduziu menos vezes, apenas três, o número de atributos nos 11 conjuntos de dados. Os algoritmos que apresentaram maior número de vezes em que houve redução no subconjunto de atributos selecionados foram as duas versões de FDimBF, as quais sempre reduziram os subconjuntos de atributos selecionados. Por outro lado, esses dois algoritmos apresentaram degradação da performance do modelo de classificação induzido, com diferença estatisticamente significativa, em sete do total de 77 comparações entre as médias dos erros dos modelos induzidos considerando os subconjuntos de atributos selecionados pelos algoritmos de seleção de atributos.

Uma análise análoga à apresentada anteriormente foi realizada comparando-se todos os algoritmos entre si. Na Tabela 8.10 são apresentados, para cada conjunto de dados (sigla e número original de atributos) e cada comparação, o número de atributos selecionados

pelos algoritmos identificados na primeira coluna. Assim como na Tabela 8.9, nessa tabela as comparações entre as médias de erros que apresentaram diferença estatisticamente significativa estão apresentadas em negrito. Na Tabela 8.11 são apresentados, para cada conjunto de dados, o número de vezes em que cada algoritmo selecionou um subconjunto menor de atributos considerando as comparações em que os algoritmos apresentaram estatisticamente performances similares em relação ao erro. Por exemplo, para o conjunto de dados Breast Cancer (BCa), o algoritmo $\mathcal{C}4.5$ comparado a todos os outros algoritmos, selecionou três vezes ($\mathcal{C}4.5$-ReliefF, $\mathcal{C}4.5$-CFS e $\mathcal{C}4.5$-FCBF) subconjuntos de atributos menores (segunda coluna e terceira linha da Tabela 8.10).

	BCa 9	Bup 6	Ger 24	Hun 10	Ion 34	Pim 8	Sat 36	Seg 19	Son 60	Veh 18	Wav 21
$\mathcal{C}4.5$-ReliefF	7–9	6–5	24–24	9–10	15–33	7–8	36–36	16–18	15–60	18–18	21–21
$\mathcal{C}4.5$-CFS	7–9	6–1	24–2	9–3	15–14	7–3	36–23	16–5	15–19	18–11	21–15
$\mathcal{C}4.5$-FCBF	7–9	6–1	24–15	9–6	15–33	7–8	36–36	16–18	15–21	18–18	21–19
$\mathcal{C}4.5$-CBF	7–7	6–1	24–15	9–5	15–7	7–8	36–12	16–9	15–14	18–18	21–12
$\mathcal{C}4.5$-DF(1)	7–3	6–4	24–12	9–3	15–3	7–3	36–6	**16–4**	15–5	18–6	**21–3**
$\mathcal{C}4.5$-DF(2)	7–3	6–4	24–12	9–4	15–4	7–4	36–6	**16–4**	15–10	**18–6**	**21–3**
ReliefF-CFS	9–9	5–1	24–2	10–3	33-14	8–3	36–23	18–5	60–19	18–11	21–15
ReliefF-FCBF	9–9	5–1	24–15	10–6	33–33	8–8	36–36	18–18	60–21	18–18	21–19
ReliefF-CBF	9–7	5–1	24–15	10–5	33–7	8–8	36–12	18–9	60–14	18–18	21–12
ReliefF-DF(1)	9–3	5–4	24–12	10–3	33–3	8–3	36–6	**18–4**	**60–5**	**18–6**	**21–3**
ReliefF-DF(2)	9–3	5–4	24–12	10–4	**33–4**	8–4	**36–6**	**18–4**	60–10	**18–6**	**21–3**
CFS-FCBF	9–9	1–1	2–15	3–6	14–33	3–8	23–36	5–18	19–21	11–18	15–19
CFS-CBF	9–7	1–1	2–15	3–5	14–7	3–8	23–12	5–9	19–14	11–18	15–12
CFS-DF(1)	9–3	1–4	2–12	3–3	14–3	3–3	**23–6**	**5–4**	**19–5**	11–6	**15–3**
CFS-DF(2)	9–3	1–4	2–12	3–4	14–4	3–4	**23–6**	**5–4**	19–10	11–6	**15–3**
FCBF-CBF	9–7	1–1	15–15	6–5	33–7	8–8	36–12	18–9	21–14	18–18	19–12
FCBF-DF(1)	9–3	1–4	15–12	6–3	33–3	8–3	36–6	**18–4**	21–5	18–6	**19–3**
FCBF-DF(2)	9–3	1–4	15–12	6–4	**33–4**	8–4	36–6	**18–4**	21–10	18–6	**19–3**
CBF-DF(1)	7–3	1–4	15–12	5–3	7–3	8–3	**12–6**	**9–4**	14–5	18–6	12–3
CBF-DF(2)	7–3	1–4	15–12	5–4	7–4	8–4	**12–6**	**9–4**	14–10	18–6	12–3
DF(1)-DF(2)	3–3	4–4	12–12	3–4	3–4	3–4	6–6	4–4	5–10	6–6	3–3

Tabela 8.10: Comparação entre os números de atributos selecionados pelos algoritmos de SA. Comparação entre as médias de erros dos modelos construídos (em negrito resultados estatisticamente significativos)

Na Tabela 8.10 é também apresentado para cada algoritmo o número total de comparações para as quais o respectivo algoritmo foi o vencedor, *i.e.* selecionou menos atributos comparado com cada um dos outros algoritmos (Ganhos). Na última coluna dessa tabela, é informado o número total de conjuntos de dados para os quais cada algoritmo foi o vencedor (Ganhos por Conjunto de Dados). Casos nos quais foram selecionados o mesmo número de atributos por dois algoritmos não foram computados.

Do total de 11 conjuntos de dados, os algoritmos FDimBF não apresentaram boa performance para dois deles, Segment e Waveform. Para Segment, todos os subconjuntos de atributos selecionados pelas duas versões de FDimBF apresentaram erros estatisticamente maiores que os erros apresentados pelos modelos construídos utilizando os atributos selecionados pelos outros algoritmos de SA. Para Waveform, as comparações entre as duas

	BCa	Bup	Ger	Hun	Ion	Pim	Sat	Seg	Son	Veh	Wav	Ganhos	Ganhos por Conjunto de Dados
ReliefF	0	1	0	0	0	0	0	0	0	0	0	1	0
$\mathcal{C}4.5$	3	0	0	1	2	3	0	2	3	0	0	14	0
CFS	0	**4**	**6**	**5**	3	**5**	3	**4**	2	**4**	3	39	6
FCBF	0	**4**	2	2	0	0	0	0	1	0	2	11	1
CBF	3	**4**	2	2	4	0	**4**	3	4	0	**4**	30	3
DF(1)	**5**	2	4	**5**	**6**	**5**	3	0	4	**4**	1	39	5
DF(2)	**5**	2	4	4	3	4	2	0	**5**	3	1	33	2

Tabela 8.11: Resumo do número de vezes em que cada algoritmo seleciona um subconjunto menor de atributos

versões de FDimBF e os algoritmos ReliefF, $\mathcal{C}4.5$, CFS e FCBF resultaram em diferenças estatisticamente significativas com erros maiores para FDimBF. Os algoritmos CBF e FDimBF(1) e (2) apresentaram erros estatisticamente similares, tendo porém as duas versões de FDimBF selecionado apenas um quarto do total de atributos selecionados por CBF. Em outros quatro conjuntos de dados houve diferença estatisticamente significativa entre os erros de FDimBF (maiores) e os outros algoritmos (menores): Ionosphere, Satimage, Sonar e Vehicle.

Embora em $12,98\%$ das comparações os algoritmos FDimBF tenham apresentado performances piores quanto ao erro dos modelos construídos com os subconjuntos de atributos selecionados, quando comparados aos outros algoritmos de SA, considerando um panorama geral do número de ganhos, as duas versões de FDimBF juntamente com o algoritmo CFS, apresentaram os maiores números de ganhos do número de vezes em que selecionaram menos atributos com performance estatisticamente similar. Do ponto de vista de ganhos por conjunto de dados, a mesma classificação geral foi seguida, tendo CFS vencido em seis do total de 11 conjuntos de dados e FDimBF(1) vencido em cinco casos.

8.6 Características dos Conjuntos de Dados Associadas à Utilização da Dimensão Fractal como uma Medida Adequada para a Seleção de Atributos

Nesta seção é apresentada a análise da relação entre as características dos conjuntos de dados e a utilização dos algoritmos FDimBF para a seleção de atributos. Os metadados são compostos por 121 exemplos[4] descritos por nove atributos — Tabela 8.12 —, os quais não possuem valores desconhecidos nem exemplos conflitantes ou duplicados.

Essa análise foi realizada sob dois aspectos:

1. Características gerais dos conjuntos de dados e adequação da utilização da dimensão fractal como medida para a remoção de atributos redundantes e

[4]Onze conjuntos de dados × 11 comparações: sem SA, FDimBF(1) e FDimBF(2), ReliefF, CFS, FCBF e CBF.

Atributo	Nome	Descrição
a1	abordagem1	conjunto original de atributos e subconjuntos de atributos selecionados por cada um dos sete algoritmos de SA considerados neste trabalho a serem comparados com abordagem2
a2	abordagem2	algoritmos de SA considerados neste trabalho
a3	comparacao	comparação entre número de atributos selecionados pela abordagem1 e pela abordagem2
a4	diferenca	se há diferença significativa entre as médias dos erros das abordagens 1 e 2 medidos por validação cruzada com 10 partições
a5	fractal	característica fractal do conjunto de dados
a6	pontos	número de pontos utilizados para construir a curva de comportamento do conjunto de dados
a7	proporcao	proporção do número de exemplos em relação ao número de atributos
a8	formato-orig	formato da distribuição da maioria dos atributos originais do conjunto de dados sem SA

Tabela 8.12: Descrição dos atributos da metabase

2. Padrões encontrados na aplicação dos algoritmos FDimBF para os conjuntos de dados considerados neste trabalho.

Desse modo, os metadados foram organizados em duas metabases, Meta1 e Meta2, para as quais é apresentado o resumo das características na Tabela 8.13.

Conjunto de Dados	# Exemplos	Atributos Utilizados	# Atributos (num.,nom.)	Classes	Classe %	Erro da CM
Meta1	121	a5, a6, a7 e a8	4 (2,2)	excelente muito bom bom regular ruim	27,27% 49,59% 4,96% 4,96% 13,22%	50,41% sobre muito bom
Meta2	121	todos	8 (2,6)	excelente muito bom bom regular ruim	27,27% 49,59% 4,96% 4,96% 13,22%	50,41% sobre muito bom

Tabela 8.13: Resumo das metabases

É importante notar que para essas duas metabases Meta1 (características gerais dos conjuntos de dados descritas pelos atributos a5, a6, a7 e a8) e Meta2 (características gerais dos conjuntos de dados associadas às características da aplicação dos algoritmos FDimBF descritas por todos os nove atributos apresentados na Tabela 8.12), foi considerada como classe o desempenho das versões FDimBF(1) e FDimBF(2) em relação à classificação no gráfico de percentagem de atributos selecionados por média do erro do modelo construído utilizando validação cruzada com 10 partições — Seção 8.4 na página 25.

Para cada uma das metabases, foi utilizada a ferramenta $\mathcal{S}ee5$ (Rulequest-Research, 1999), uma versão posterior do algoritmo $\mathcal{C}4.5$, executado com valores default, para a indução de regras de decisão. A seguir são apresentadas essas duas análises.

8.6.1 Características Gerais dos Conjuntos de Dados e Adequação do Uso dos Algoritmos FDimBF

A primeira metabase Meta1 contém quatro atributos, a5, a6, a7 e a8 — Tabela 8.13 — os quais apresentam, como mencionado, informações sobre características gerais dos conjuntos de dados:

- a5: Característica fractal do conjunto de dados (Seção 8.1 na página 16);
- a6: Número de pontos utilizados para construir a curva de comportamento do conjunto de dados (Seção 8.1 na página 16);
- a7: Proporção de exemplos por atributos e
- a8: Formato da distribuição da maioria dos atributos originais do conjunto de dados sem SA (Seção 8.3 na página 22).

Como mencionado anteriormente, foi considerado como classe o desempenho de cada um dos casos em relação à classificação no gráfico de percentagem de atributos selecionados por média do erro do modelo construído utilizando validação cruzada com 10 partições — Seção 8.4 na página 25.

O objetivo dessa primeira análise foi verificar se existe alguma relação entre o desempenho dos algoritmos FDimBF e as características gerais dos conjuntos de dados considerados. Em outras palavras, o intuito foi encontrar características dos conjuntos de dados que pudessem prover uma idéia se a utilização da dimensão fractal como medida para a remoção de atributos redundantes era adequada e, conseqüentemente, se os algoritmos FDimBF poderiam ser apropriados para a seleção de atributos relevantes e não redundantes.

As regras induzidas por $\mathcal{S}ee5$ utilizando o Meta1 encontram-se no Apêndice E. O modelo induzido consiste de oito regras e o erro aparente desse modelo é de $4,5\%$, o qual encontra-se concentrado na classe Muito Bom. A estimativa do erro desse modelo e o erro padrão utilizando 10 partições com validação cruzada são de $6,1\%$ e $1,5\%$, respectivamente.

Considerando como critério de importância o número de vezes em que os atributos aparecem nas regras induzidas, os atributos mais importantes para essa metabase, foram a proporção de exemplos por atributo (em seis das oito regras) e o formato da distribuição da maioria dos atributos do conjunto original de dados (em cinco das nove regras).

Em relação à proporção de exemplos por atributo na representação de conjuntos de dados, não há um consenso sobre que proporção seria adequada, porém, uma regra geral é que quanto maior essa proporção melhor deve ser essa representação. A determinação dessa quantidade depende de diversos fatores, dentre eles os métodos que serão usados para explorar e construir modelos a partir desses dados e a própria complexidade do domínio ao qual esses dados pertencem. Dentre as proporções recomendadas na literatura, há

a descrição de que uma proporção mínima aceitável seria de cinco exemplos para cada atributo. Porém há autores que defendam que uma proporção mais aceitável seria de 10 exemplos, enquanto outros propõe que essa proporção deva ser de 20 exemplos por atributo (Hair et al., 1998).

As regras induzidas com o conjunto de metadados Meta1 mostraram diversos padrões consistentes com o conhecimento prévio, tal qual a regra:

```
SE proporcao > 96,13
   ENTÃO classe = Muito Bom [36; 0,974]
```

Os números entre colchetes indicam que essa regra cobre 36 exemplos do total de 78 exemplos dessa classe do conjunto de metadados Meta1 com grau de confiança de 0,974. Essa regra indica que havendo uma alta proporção entre o número de exemplos por atributo, o conjunto de dados apresentaria uma performance muito boa utilizando o algoritmo FDimBF.

Outra regra que apresentou conhecimento consistente tem como base a proporção de exemplos por atributo e o número de pontos considerados para construir a curva de comportamento do conjunto de dados, mostrada a seguir:

```
SE pontos <= 6
   E proporcao > 10,32
   E proporcao <= 41,67
   ENTÃO classe = Excelente [18; 0,950]
```

Uma análise dos exemplos cobertos por essa regra, revelou que todos apresentavam entre cinco e seis pontos considerados para construir a curva de comportamento do conjunto de dados e proporção de exemplos por atributo variando de 26,1 a 41,67. Assim, conjuntos de dados apresentando razoável número de pontos e proporção de exemplos por atributo, teriam boa probabilidade de apresentarem excelente performance utilizando o algoritmo FDimBF.

Outra regra interessante, apresentada a seguir, indica que conjuntos de dados apresentando formato da distribuição Tipo 2 (T_2) para a maioria dos atributos do conjunto original de dados permitiriam a seleção de subconjuntos de atributos, por meio do algoritmo FDimBF, que gerariam muito bons modelos.

```
SE formato-orig = Tipo 2
   ENTÃO classe = Muito Bom [12; 0,929]
```

É interessante ressaltar que os exemplos classificados como Excelente apresentam formato da distribuição da maioria dos atributos do conjunto original de dados do Tipo 1 (T_1) e os exemplos classificados como Muito Bom foram todos classificados como dos Tipos 1 e 2.

8.6.2 Padrões Encontrados na Aplicação dos Algoritmos FDimBF para os Conjuntos de Dados

Como citado anteriormente, a segunda metabase Meta2 — Tabela 8.13 na página 32 — contém informações sobre características gerais dos conjuntos de dados e características da aplicação dos algoritmos FDimBF. Essas informações são descritas por todos os oito atributos apresentados na Tabela 8.12 e os exemplos classificados do mesmo modo como foram rotulados em Meta1.

O modelo induzido com essa metabase consiste de 12 regras. O erro aparente desse modelo é de $0,0\%$ e a estimativa do erro e o erro padrão utilizando 10 partições com validação cruzada são ambos de $0,0\%$.

Nesse conjunto de metadados, os atributos considerados mais importantes, segundo o critério de importância de número de vezes em que os atributos aparecem nas regras induzidas, foram o número de pontos utilizados para construir a curva de comportamento do conjunto de dados (nove vezes), juntamente com a proporção de exemplos por atributo (oito vezes) e o formato da distribuição da maioria dos atributos do conjunto original de dados (sete vezes). A determinação desses atributos como sendo os mais importantes está em sintonia com o conhecimento prévio.

Dentre essas 12 regras algumas apresentaram conhecimento similar ao apresentado pelas regras induzidas usando Meta1. Entre elas, a de que conjuntos de dados apresentando formato da distribuição Tipo 2 (T_2) para a maioria dos atributos do conjunto original de dados permitiriam a seleção de subconjuntos de atributos, por meio do algoritmo FDimBF, que gerariam muito bons modelos.

Outra regra interessante, a qual também considera o formato da distribuição da maioria dos atributos do conjunto original de dados, é a seguinte:

```
SE diferenca = nao
   E formato-orig = Tipo 1
   ENTÃO classe = Excelente [24; 0,962]
```

Nessa regra, a qual cobre 24 exemplos, o algoritmo FDimBF apresentando performance estatisticamente similar aos outros algoritmos de seleção de atributos considerados, e o conjunto de dados apresentando formato da distribuição Tipo 1 (T_1), permitiriam a seleção de subconjuntos de atributos para a construção de modelos com performances excelentes.

Um outra regra interessante é a seguinte:

```
SE abordagem2 = FDimBF(1)
   E pontos <= 5
   E proporcao > 26,1
   E formato = Tipo 3
   ENTÃO classe = Muito Bom [18; 0,950]
```

pois cobre bem muitos exemplos. Entretanto, o formato de distribuição do Tipo 3 (T_3) também participa de outra regra que que cobre poucos exemplos, utiliza FDimBF(2) e cuja classe é Ruim.

Adicionalmente, foram encontrados outros padrões que descrevem relações entre a aplicação dos algoritmos, o número de pontos da curva de comportamento, a característica fractal e a distribuição dos valores da maioria dos atributos originais do conjunto de dados. Uma das regras que descreve esses padrões indica que mesmo que a característica fractal do conjunto de dados seja mediana, se o número de pontos não for pequeno, o desempenho será também muito bom (cobertura de 24 exemplos com grau de confiança de 0,962).

```
SE fractal = regular
   E pontos > 5
   ENTÃO classe = Muito Bom [24; 0,962]
```

9 Considerações Finais

Neste trabalho foi apresentada a proposta de um algoritmo, FDimBF, para a seleção de atributos importantes, bem como uma série de experimentos, nos quais a abordagem proposta é comparada à alguns algoritmos freqüentemente citados na literatura. FDimBF realiza a SA em duas etapas: seleção de atributos relevantes em relação à classe e remoção de atributos redundantes. A primeira etapa é realizada utilizando tanto uma medida baseada em ganho de informação — FDimBF(1) — quanto uma medida baseada em distância — FDimBF(2). A segunda etapa é realizada utilizando como medida de correlação a dimensão fractal do conjunto de dados. Ao final desse processo, o algoritmo terá selecionado um subconjunto de atributos relevantes e não redundantes. Deve ser observado que a maioria dos algoritmos de SA não trata esses dois problemas, pois considera como atributos importantes somente aqueles que são relevantes em relação à classe.

Os resultados obtidos utilizando 11 conjuntos de dados e outros algoritmos de SA mostram que o algoritmo proposto é comparável a outros algoritmos de SA, selecionando os menores subconjuntos de atributos importantes com performances similares a algoritmos como o CFS (*Correlation-Based Feature Selection*). Assim, consideramos que a DF pode ser também considerada uma boa candidata para realizar seleção de atributos na área de aprendizado de máquina, na qual não é de nosso conhecimento que ela tenha sido utilizada.

Referências

Batista, G. E., R. C. Prati, and M. C. Monard (2004). A study of the behavior of several methods for balancing machine learning data. *SIGKDD Explorations: Special issue on Learning from Imbalanced Datasets 6*(1), 20–29. 7

Batista, G. E. A. P. A. (2001). Sintaxe padrão do arquivo de exemplos do projeto DISCOVER. http://www.icmc.sc.usp.br/~gbatista/SintaxePadraoFinal.htm. 15

Batista, G. E. A. P. A. (2003). Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. Tese de Doutorado, ICMC-USP, http://www.icmc.usp.br/~gbatista/pdfs/TeseDoutorado.pdf. 15

Batista, G. E. A. P. A. and M. C. Monard (2002). A Study of K-Nearest Neighbour as an Imputation Method. In A. Abraham, J. R. del Solar, and M. Köppen (Eds.), *Soft Computing Systems: Design, Management and Applications*, Santiago, Chile, pp. 251–260. IOS Press. http://www.icmc.usp.br/~gbatista. 15

Batista, G. E. A. P. A. and M. C. Monard (2003a). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence 17*(5), 519–533. 14

Batista, G. E. A. P. A. and M. C. Monard (2003b). Descrição da Arquitetura e do Projeto do Ambiente Computacional DISCOVER LEARNING ENVIRONMENT — DLE. Technical Report 187, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_187.pdf. 15

Batista, G. E. A. P. A. and M. C. Monard (2005). The DISCOVER OBJECT LIBRARY — DOL user's manual. Technical report, ICMC-USP. (em preparação). 15

Brassard, G. and P. Bratley (1997). *Fundamentals of Algorithms*. New Jersey: Prentice Hall. 11

Faloutsos, C. and I. Kamel (1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *Proc.of the 13th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Minneapolis, MN, pp. 4–13. 4, 5

Fayyad, U. M. and K. B. Irani (1993). Multi-interval discretisation of continuous-valued attributes. In *Proc. of the Thirteenth International Joint Conference on Artificial Intelligence*, pp. 1022–1027. Morgan Kaufmann. 10, 11

Hair, P. E., R. Anderson, R. Tatham, and W. Black (1998). *Multivariate Data Analysis*. New Jersey: Prentice Hall. 34

Hall, M. (1999). *Correlation-based Feature Subset Selection for Machine Learning*. Ph. D. thesis, Department of Computer Science. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf. 12

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proc. of the 17th Int. Conf. on Machine Learning*, San Francisco, CA, pp. 359–366. Morgan Kaufmann. 2, 11

Kira, K. and L. Rendell (1992). A pratical approach to feature selection. In *Proc. of the 9th Int. Conf. on Machine Learning*, Aberdeen, Scotland, pp. 249–256. Morgan Kaufmann. 9

Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artif. Intell. 97*(1-2), 273–324. 2

Koller, D. and M. Sahami (1996). Toward optimal feature selection. In *Proc. of the 13th Int. Conf. on Machine Learning*, Bari, Italy, pp. 284–292. 2

Kononenko, I. (1994). Estimating attributes: Analysis and extention of Relief. Amsterdam, pp. 171–182. 10

Lee, H. D. and M. C. Monard (2003). Seleção de atributos para algoritmos de aprendizado de máquina supervisionado utilizando como filtro a dimensão fractal. *Revista de La Sociedad Chilena de Ciencia de La Computación 4*(1), 1–8. 5, 12

Lee, H. D., M. C. Monard, R. F. Voltolini, and F. C. Wu (2005). Proposta de um algoritmo de seleção de atributos importantes para aprendizado supervisionado utilizando a dimensão fractal para tratamento de redundância: Avaliação experimental. In *Proc. of the Sixth Workshop on Artificial Intelligence, Jornadas Chilenas de Computación*, Valdivia, Chile. (in print). 5, 12

Lee, H. D., M. C. Monard, and F. C. Wu (2005a). Feature subset selection for supervised learning using fractal dimension. In *Proc.of the Fifth Congress of Logic Applied to Technology*, Himeji, Japan. (in print). 5, 12

Lee, H. D., M. C. Monard, and F. C. Wu (2005b). Seleção de atributos relevantes e não redundantes usando a dimensão fractal do conjunto de dados. In *Anais do V Encontro Nacional de Inteligência, XXV Congresso da Sociedade Brasileira de Computação*, Porto Alegre, RS, pp. 444–453. 5, 12

Liu, H. and H. Motoda (1998). *Feature Selection for Knowledge and Data Mining*. Massachusetts: Kluwer Academic Publishers. 1

Liu, H. and R. Setiono (1996). A probabilistic approach to feature selection – a filter solution. In *Proc.of the 13th Int. Conf. on Machine Learning*, Bari, Italy, pp. 319–327. 11

Mandelbrot, B. B. (1985). *The Fractal Geometry of Nature: Updated and Augmented*. New York: W. H. Freeman and Company. 3

Michie, D., D. J. Spiegelhalter, C. C. Taylor, and J. Campbell (Eds.) (1994). *Machine learning, neural and statistical classification*. New Jersey: Ellis Horwood. 7

Motulsky, H. (1995). *Intuitive Biostatistics*. New York: Oxford University Press. 16

Newman, D., S. Hettich, C. Blake, and C. Merz (1998). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`. 7

Prati, R. C., J. A. Baranauskas, and M. C. Monard (2001a). Extração de informações padronizadas para a avaliação de regras induzidas por algoritmos de aprendizado de máquina simbólico. Technical report, ICMC-USP. `ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_145.ps.zip`. 15

Prati, R. C., J. A. Baranauskas, and M. C. Monard (2001b). Uma proposta de unificação da linguagem de representação de conceitos de algoritmos de aprendizado de máquina simbólicos. Technical Report 137, ICMC-USP. ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_137.ps.zip. 15

Prati, R. C., J. A. Baranauskas, and M. C. Monard (2002). Padronização da Sintaxe e Informações sobre Regras Induzidas a Partir de Algoritmos de Aprendizado de Máquina Simbólico. *Revista Eletrônica de Iniciação Científica 2*(3). http://www.sbc.org.br/reic/edicoes/2002e3. 15

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing*. New York: Cambridge University Press. 5, 10, 11

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn. 1*(1), 81–106. 12

Quinlan, J. R. (1993). $\mathcal{C}4.5$*: Programs for Machine Learning*. Morgan Kaufmann. California. 6, 9, 14

Robnik-Sikonja, M. and I. Kononenko (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn. 53*(1-2), 23–69. 6, 10

Rulequest-Research (1999). Data mining tools $\mathcal{S}ee5$ and $\mathcal{C}5.0$. http://www.rulequest.com/see5-info.html. 32

Sousa, E. P. M., C. Traina, A. J. M. Traina, and C. Faloutsos (2002). How to use fractal dimension to find correlations between attributes. In *Workshop Notes of KDD 2002 Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches*, Edmonton, Canada, pp. 26–30. 4

Traina, C., A. J. M. Traina, and C. Faloutsos (2003). MDE – measure distance exponent manual. (Internal Document). 6

Traina, C., A. J. M. Traina, L. Wu, and C. Faloutsos (2000). Fast feature selection using fractal dimension. In *Proc. of the 15th Brazilian Data Base Symposium*, João Pessoa, Brasil, pp. 158–171. 4, 5, 6

Witten, I. H. and E. Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. California: Morgan Kaufmann. 12, 14

Yu, L. and H. Liu (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research 5*, 1205–1224. iii, 1, 3, 5, 10, 11

A Dimensão Fractal e Comportamento Fractal dos Conjuntos de Dados

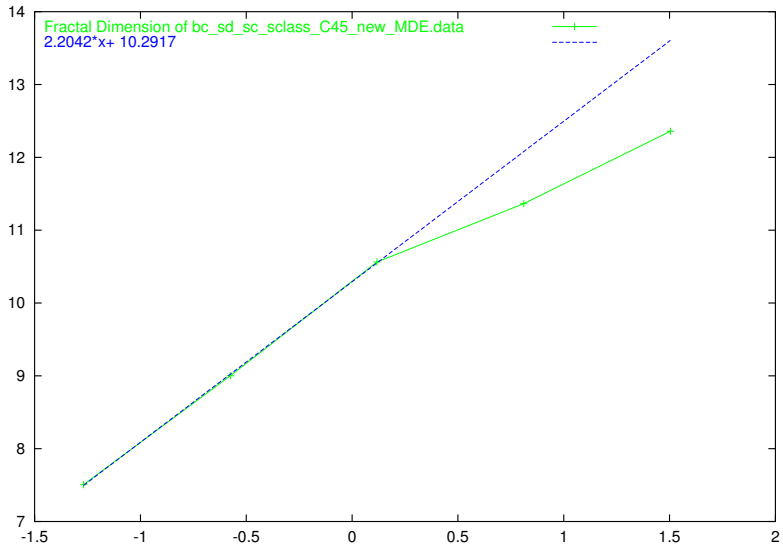


Figura A.1: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Breast Cancer

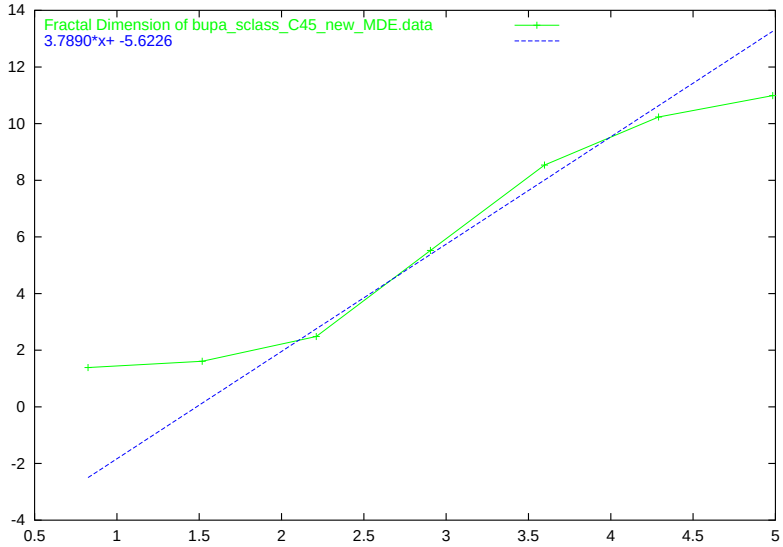


Figura A.2: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Bupa

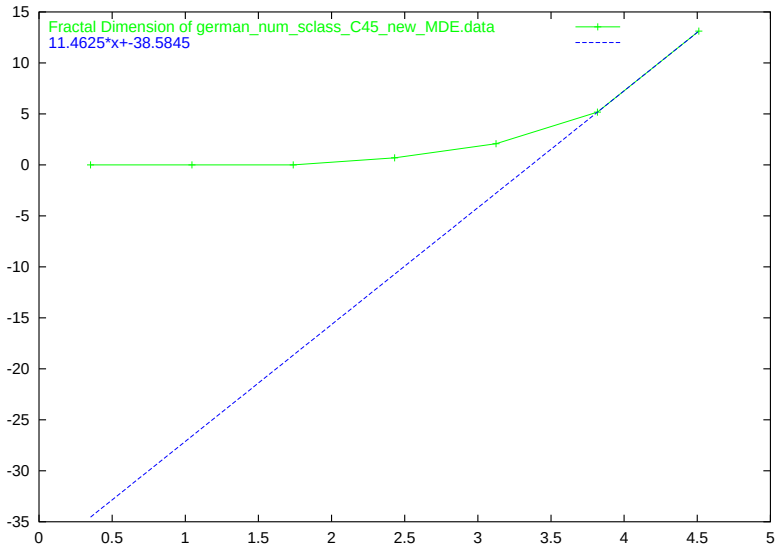


Figura A.3: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - German

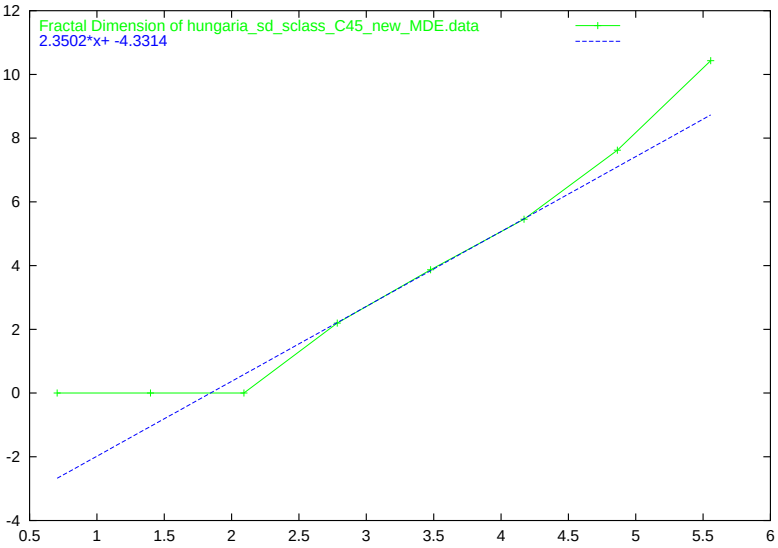


Figura A.4: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Hungarian

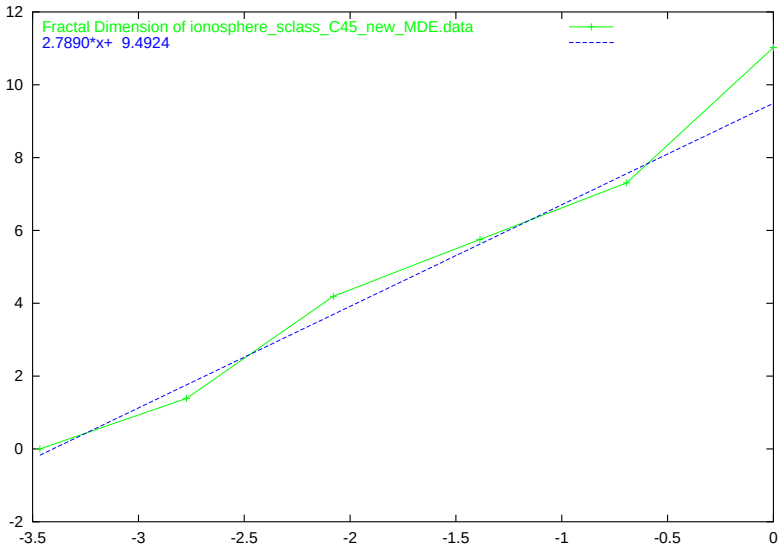


Figura A.5: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Ionosphere

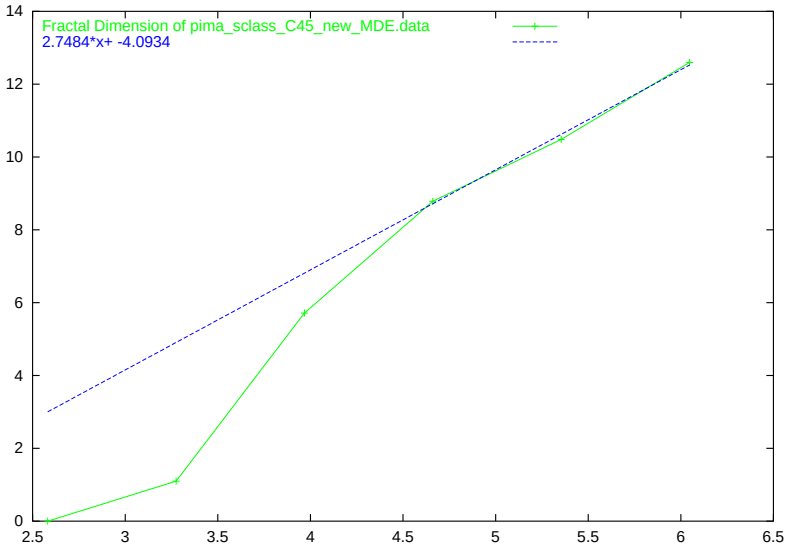


Figura A.6: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Pima

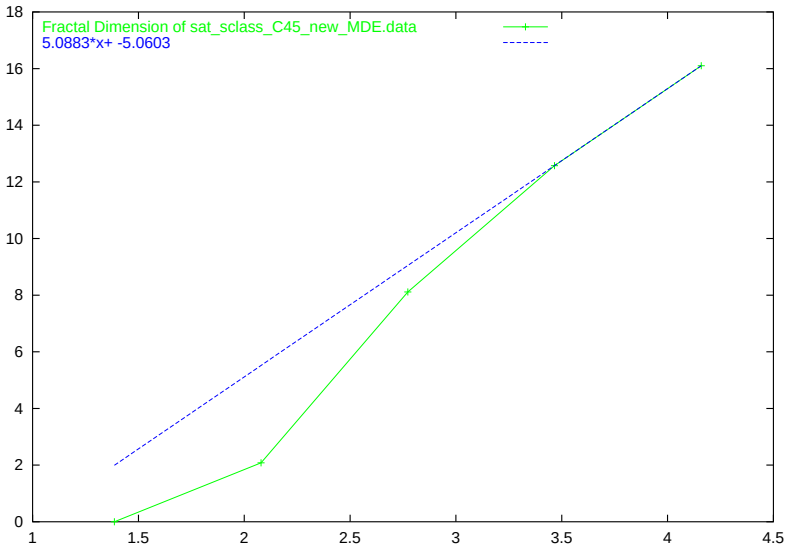


Figura A.7: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Satimage

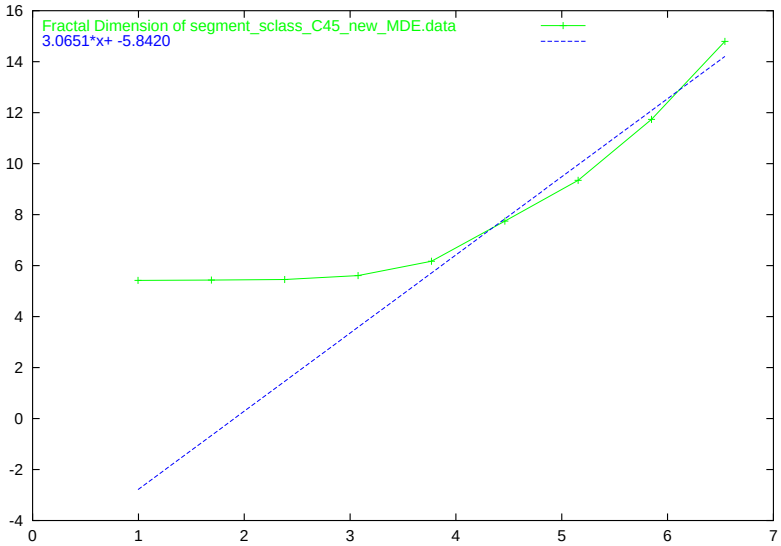


Figura A.8: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Segment

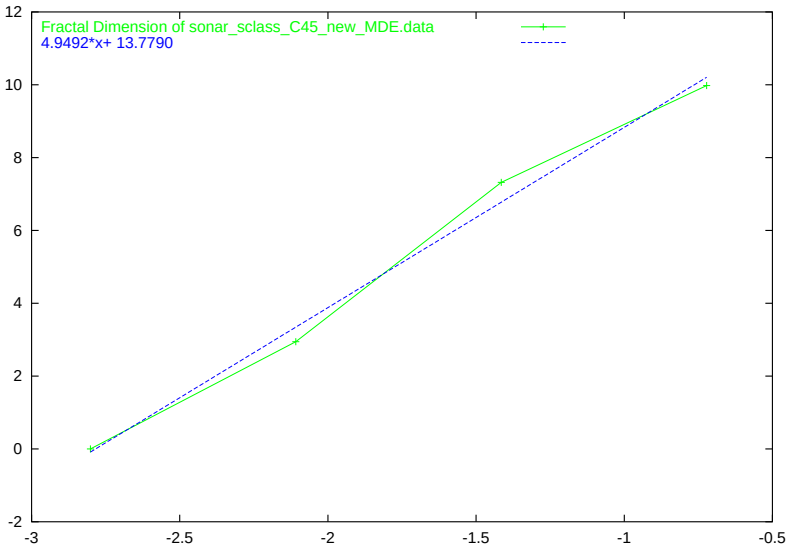


Figura A.9: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Sonar

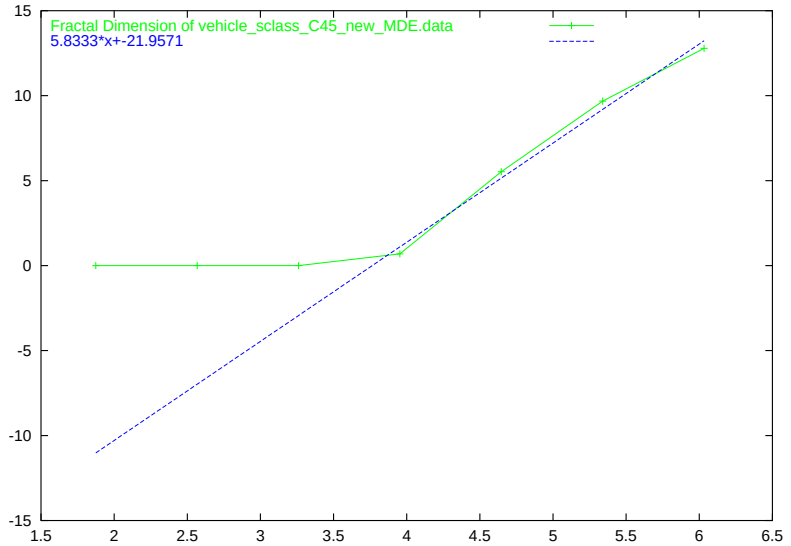


Figura A.10: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Vehicle

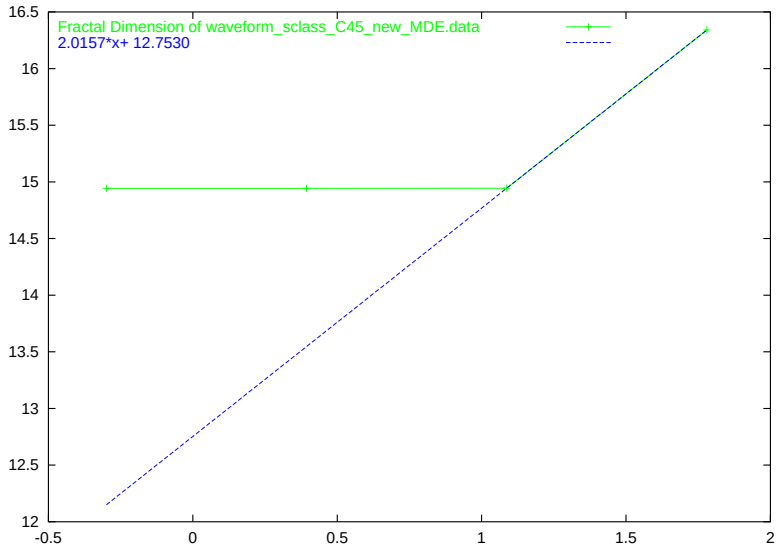


Figura A.11: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(1) - Waveform

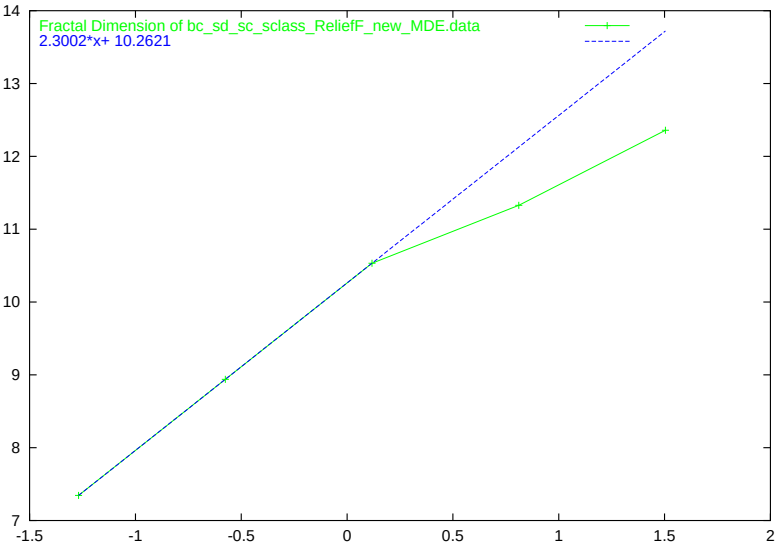


Figura A.12: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Breast Cancer

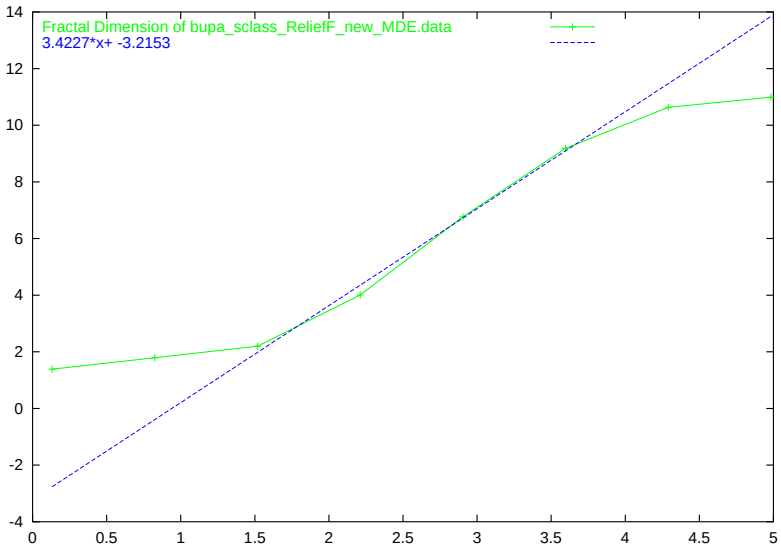


Figura A.13: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Bupa

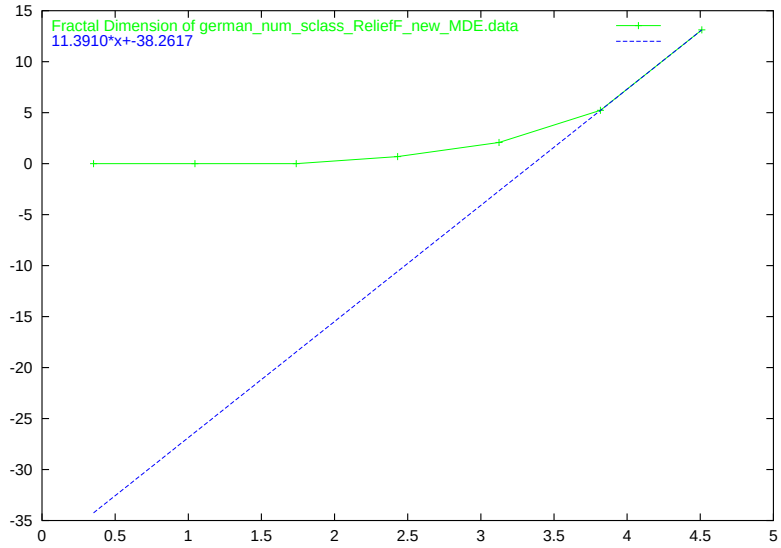


Figura A.14: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - German

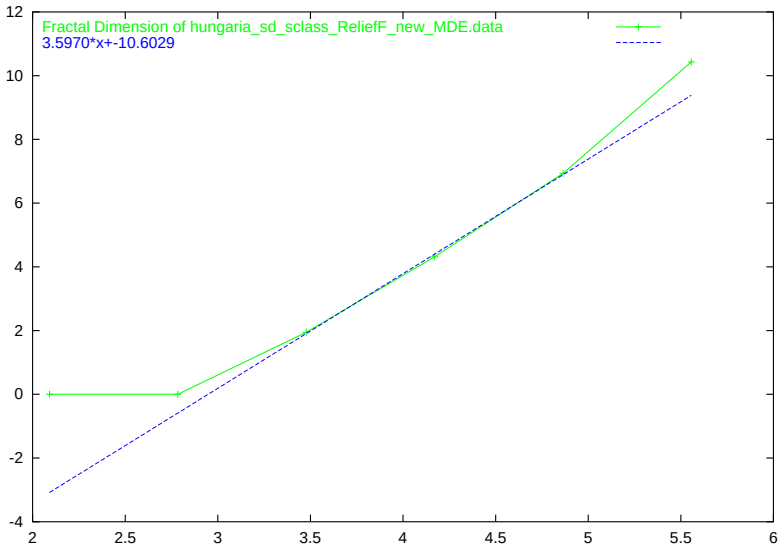


Figura A.15: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Hungarian

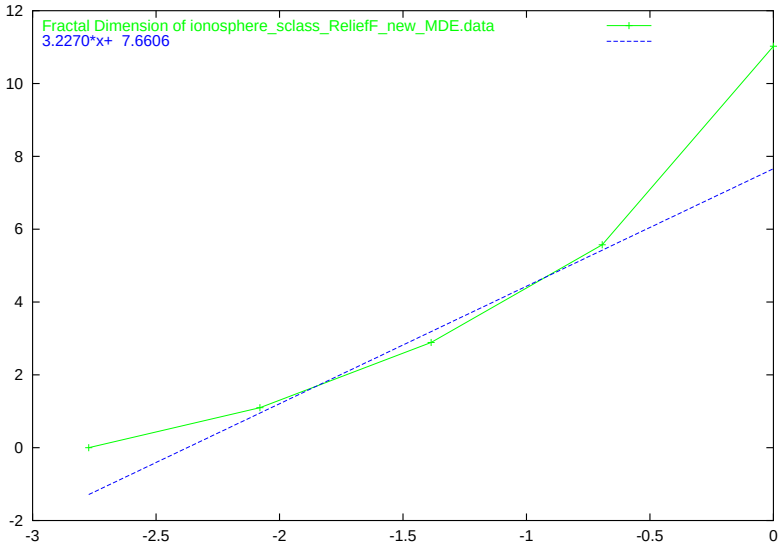


Figura A.16: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Ionosphere

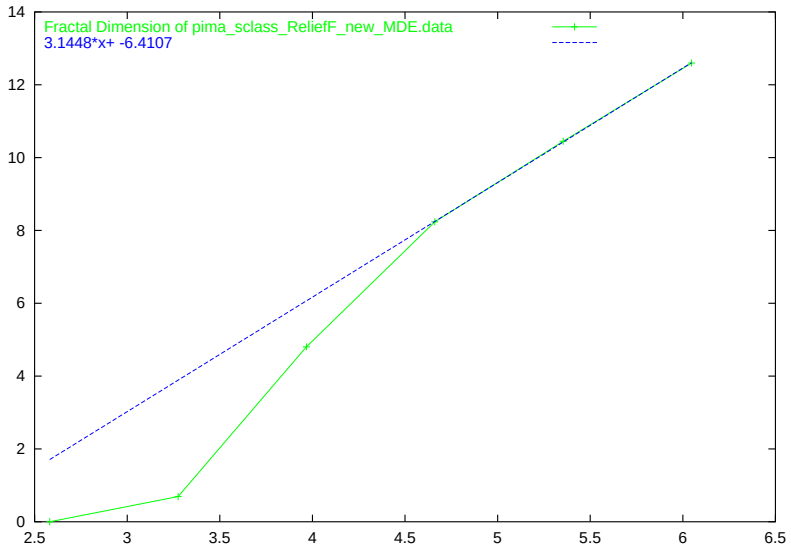


Figura A.17: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Pima

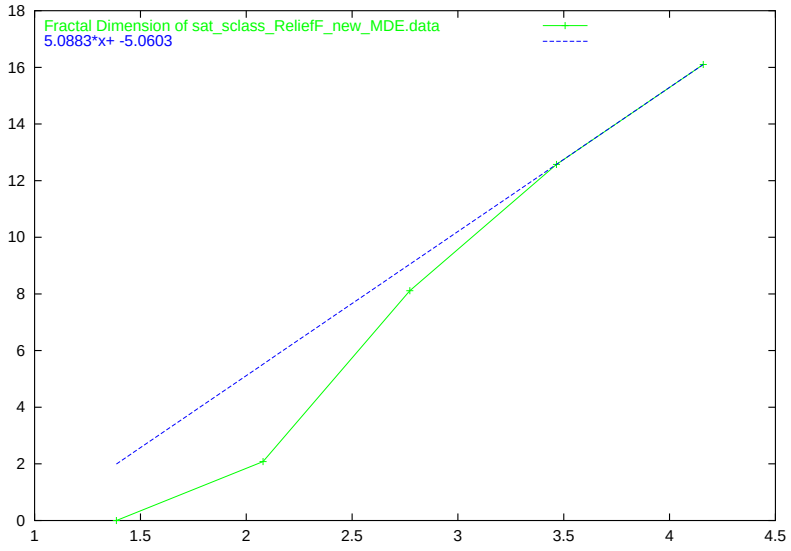


Figura A.18: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Satimage

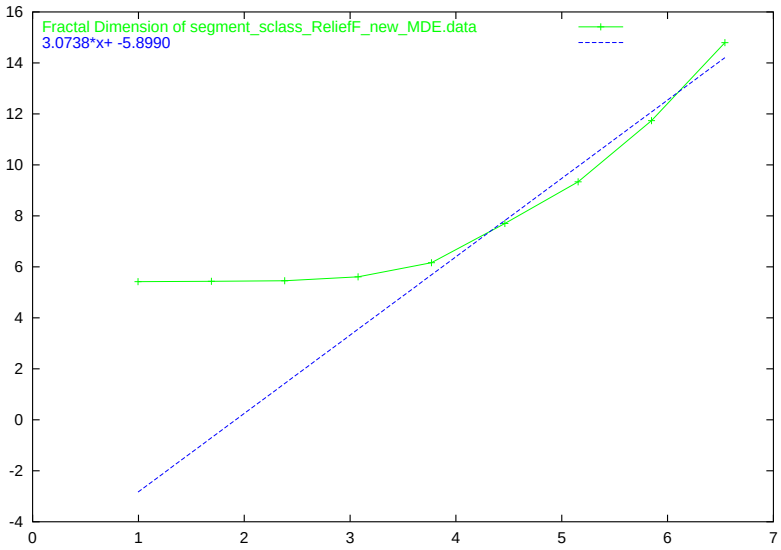


Figura A.19: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Segment

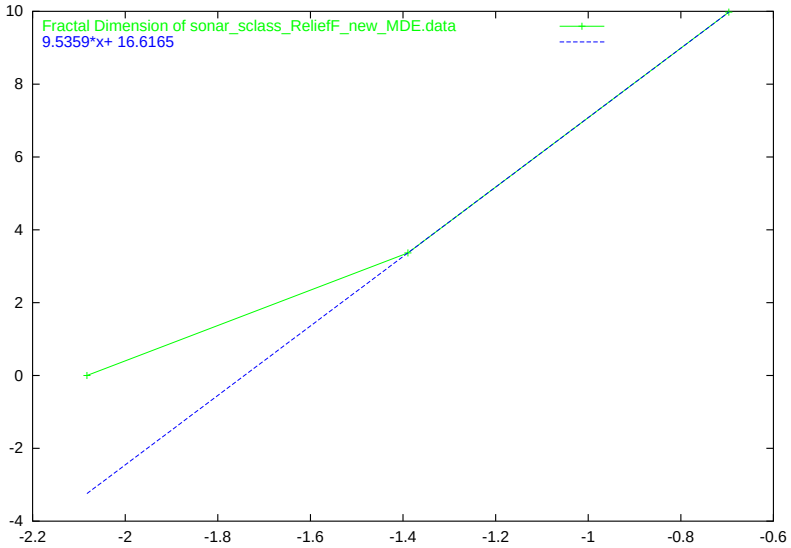


Figura A.20: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Sonar

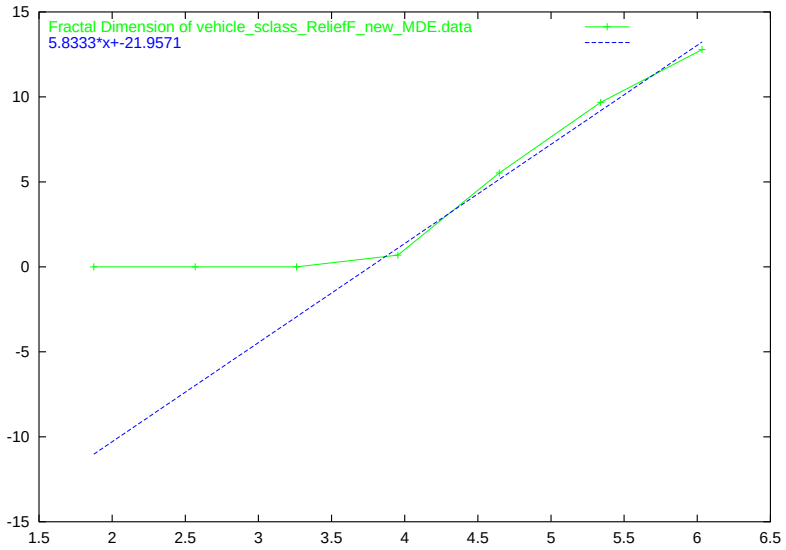


Figura A.21: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Vehicle

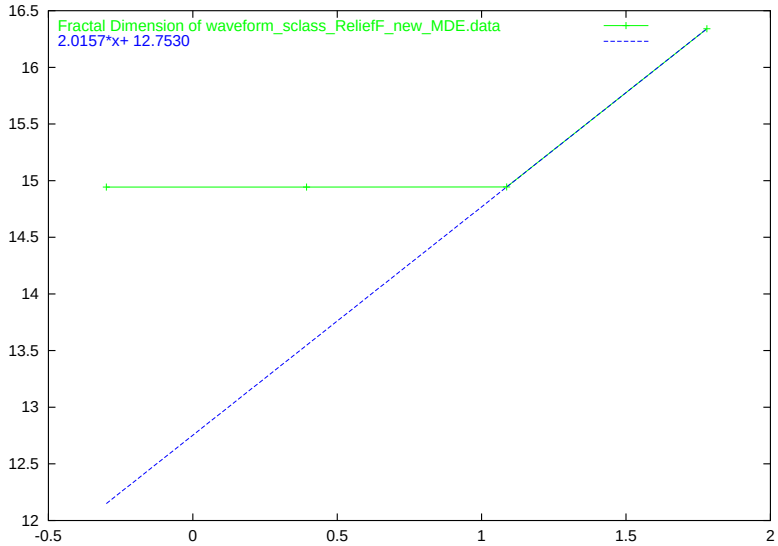


Figura A.22: Gráfico gerado utilizando o método *Box Count Plot* para FDimBF(2) - Waveform

B Subconjuntos de Atributos Selecionados

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
Clump Thickness		*** 4**	*** 2**	*	*	*		
Uniformity of Cell Size		*** 1**	*** 4**	*	*	*		
Uniformity of Cell Shape		*** 3**	*** 3**	*	*		*** 2**	*** 2**
Marginal Adhesion		*** 5**	*** 7**	*	*	*		
Single Epithelial Cell Size		*** 7**	*** 8**	*	*	*	*** 3**	*** 3**
Bare Nuclei		*** 2**	*** 1**	*	*	*		
Bland Chromatin		*** 6**	*** 6**	*	*	*	*** 1**	*** 1**
Normal Nucleoli			*** 5**	*	*	*		
Mitoses			*** 9**	*	*			
classe								
# Atrib.	9	7	9	9	9	7	3	3
% Atrib.		77,78	100,00	100,00	100,00	77,78	33,33	33,33
Média	4,54	4,83	4,54	4,54	4,54	4,98	4,40	4,40
Erro Padrão	0,70	0,54	0,70	0,70	0,70	0,62	0,54	0,54

Tabela B.1: Atributos Selecionados - Breast Cancer

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
mcv		*** 6**	*** 5**				*** 2**	*** 2**
alkphos		*** 5**					*** 1**	
sgpt		*** 3**	*** 1**					*** 1**
sgot		*** 4**	*** 4**				*** 4**	*** 3**
gammagt		*** 1**	*** 2**	*	*	*		*** 4**
drinks		*** 2**	*** 3**				*** 3**	
classe								
# Atrib.	6	6	5	1	1	1	4	4
% Atrib.		100,00	83,33	16,67	16,67	16,67	66,67	83,33
Média	29,57	32,47	33,63	36,77	36,77	36,77	42,01	33,03
Erro Padrão	2,38	2,58	3,11	2,72	2,72	2,72	1,37	2,17

Tabela B.2: Atributos Selecionados – Bupa

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
1		*** 3**	*** 51**				*** 2**	
2		*** 15**	*** 49**					
3			*** 52**					
4		*** 4**	*** 50**	*	*	*		
5			*** 47**	*	*	*		
6			*** 40**					
7			*** 60**					
8		*** 11**	*** 20**				*** 4**	
9			*** 5**	*	*	*		
10			*** 3**	*	*	*		
11		*** 1**	*** 2**	*	*	*		
12			*** 1**	*	*			
13			*** 8**	*	*	*		
14			*** 42**					
15			*** 38**					
16			*** 30**					

continua na próxima página

Tabela B.9: Atributos Selecionados – Sonar (continuação)

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
17			*** 24**					
18			*** 46**					*** 9**
19			*** 48**					
20			*** 22**		*	*		
21		*** 9**	*** 15**	*	*	*		
22			*** 32**					*** 10**
23		*** 14**	**35**				*** 1**	
24			*** 39**					
25			*** 31**					
26			*** 28**					
27		*** 2**	*** 29**					
28		*** 6**	*** 17**	*	*			
29			*** 21**					
30			**25**					
31			*** 14**					
32			*** 18**					
33			*** 33**					*** 8**
34			*** 16**					*** 7**
35			*** 19**		*			*** 6**
36			*** 4**	*	*	*		*** 5**
37			*** 11**					*** 4**
38			*** 36**					*** 3**
39			*** 34**					*** 2**
40			*** 41**					*** 1**
41			*** 27**					
42			*** 26**					
43			*** 23**					
44			*** 13**	*	*			
45		*** 10**	*** 6**	*	*	*	*** 5**	
46			*** 10**	*	*	*		
47			*** 12**	*	*	*		
48			*** 7**	*	*	*		
49			*** 9**	*	*			
50			*** 54**					
51		*** 12**	*** 43**	*	*			
52			*** 37**	*	*	*		
53		*** 7**	*** 45**					
54		*** 5**	*** 44**	*	*		*** 3**	
55		*** 13**	*** 57**					
56			*** 56**					
57			*** 59**					
58			*** 55**					
59			*** 58**					
60		*** 8**	*** 53**					
classe								
# Atrib.	60	15	60	19	21	14	5	10
% Atrib.		25,00	100,00	31,67	35,00	23,33	8,33	16,67
Média	22,12	24,95	22,12	23,95	26,38	25,52	38,02	34,55
Erro Padrão	2,79	2,97	2,79	2,64	1,71	4,27	2,37	3,42

Tabela B.9: Atributos Selecionados – Sonar

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
1		*** 1**	*** 1**	*	*	*	*** 1**	*** 1**
2		*** 2**	*** 9**		*	*	*** 12**	*** 12**
3		*** 3**	*** 3**	*	*	*	*** 6**	*** 6**
4		*** 4**	*** 16**		*	*		
5		*** 5**	*** 2**		*	*	*** 10**	*** 10**
6		*** 9**	*** 7**		*	*	*** 8**	*** 8**
7		*** 8**	*** 10**		*	*	*** 5**	*** 5**
8		*** 20**	*** 6**				*** 3**	*** 3**
9		*** 16**	*** 8**		*	*	*** 2**	*** 2**
10		*** 23**	*** 13**		*	*		
11		*** 10**	*** 14**		*	*		
12		*** 17**	**17**					
13		*** 14**	*** 12**					
14		*** 19**	*** 4**				*** 7**	*** 7**
15		*** 24**	*** 21**		*	*		
16		*** 7**	*** 11**		*	*	*** 11**	*** 11**
17		*** 6**	*** 5**		*	*		
18		*** 11**	*** 15**					
19		*** 18**	*** 24**					
20		*** 21**	*** 20**		*	*		*** 9**
21		*** 13**	*** 19**		*	*	*** 9**	*** 4**
22		*** 22**	*** 23**					
23		*** 15**	*** 22**					
24		*** 12**	*** 18**				*** 4**	
classe								
# Atrib.	24	24	24	2	15	15	12	12
% Atrib.		100,00	100,00	8,33	62,50	62,50	50,00	50,00
Média	25,30	25,30	25,30	28,00	26,40	26,40	25,50	26,30
Erro Padrão	1,07	1,07	1,07	0,68	1,90	1,90	1,49	0,79

Tabela B.3: Atributos Selecionados – German

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
age			*** 6**					*** 1**
sex		*** 3**	*** 2**		*	*		
cp		*** 2**	*** 1**	*	*	*	*** 3**	
trestbps		*** 9**	*** 8**					
chol		*** 8**	*** 10**					*** 3**
fbs		*** 7**	*** 7**		*		*** 1**	
restecg		*** 6**	*** 4**					
thalach		*** 5**	*** 9**		*	*	*** 2**	*** 2**
exang		*** 1**	*** 5**	*	*	*		
oldpeak		*** 4**	*** 3**	*	*	*		*** 4**
classe								
# Atrib.	10	9	10	3	6	5	4	4
% Atrib.		90,00	100,00	30,00	60,00	50,00	40,00	40,00
Média	23,40	21,87	23,40	21,48	23,38	23,40	24,53	22,21
Erro Padrão	2,05	2,01	2,05	2,89	2,78	1,97	2,45	2,18

Tabela B.4: Atributos Selecionados – Hungarian

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
1		*** 2**	*** 33**	*	*			
2								
3		*** 4**	*** 2**	*	*			
4		*** 13**	*** 27**	*	*		*** 2**	
5		*** 1**	*** 4**	*	*	*		
6		*** 8**	*** 12**	*	*	*		
7		*** 14**	*** 6**	*	*			
8		*** 3**	*** 3**	*	*	*	*** 3**	
9			*** 10**		*			*** 2**
10			*** 25**		*			
11			*** 31**		*			
12			*** 11**		*			
13			*** 20**		*	*		
14			*** 5**	*	*			
15			*** 14**		*			
16		*** 5**	*** 7**		*			
17		*** 9**	*** 28**		*			
18			*** 24**	*	*			
19		*** 7**	*** 13**		*			
20			*** 30**		*			
21		*** 6**	*** 19**	*	*			
22			*** 18**		*	*		*** 3**
23		*** 10**	*** 32**		*			
24			*** 1**		*			*** 4**
25			*** 15**		*			
26			*** 26**		*			
27		*** 15**	*** 16**	*	*	*		
28		*** 11**	*** 17**	*	*		*** 1**	
29			*** 9**	*	*			
30			*** 29**		*			
31		*** 12**	*** 23**		*			
32			*** 22**		*			
33			*** 21**		*			
34			*** 8**	*	*	*		*** 1**
classe								
# Atrib.	34	15	33	14	33	7	3	4
% Atrib.		44,12	97,06	41,18	97,06	20,59	8,82	11,76
Média	9,97	11,40	10,55	10,27	10,55	11,40	19,38	19,36
Erro Padrão	1,96	0,85	2,00	0,98	2,00	1,86	2,48	1,72

Tabela B.5: Atributos Selecionados – Ionosphere

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
number		* **5**	* **4**		*	*		* **4**
plasma		* **1**	* **1**	*	*	*	* **1**	
diastolic		* **6**	* **7**		*	*	* **3**	* **3**
triceps			* **3**		*	*		
two		* **7**	* **8**		*	*		
body		* **2**	* **2**	*	*	*		* **2**
diabetes		* **4**	* **6**		*	*		* **1**
age		* **3**	* **5**	*	*	*	* **2**	
classe								
# Atrib.	8	7	8	3	8	8	3	4
% Atrib.		87,50	100,00	37,50	100,00	100,00	37,50	50,00
Média	24,32	25,10	24,32	25,35	24,32	24,32	25,50	34,89
Erro Padrão	1,28	1,50	1,28	1,14	1,28	1,28	1,49	3,74

Tabela B.6: Atributos Selecionados – Pima

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
1		*** 11**	*** 8**	*	*	*	*** 5**	*** 5**
2		*** 28**	*** 16**		*	*		
3		*** 9**	*** 21**		*			
4		*** 10**	*** 28**	*	*			
5		*** 36**	*** 7**	*	*			
6		*** 14**	*** 19**	*	*			
7		*** 29**	*** 36**		*	*		
8		*** 20**	*** 34**		*			
9		*** 17**	*** 9**	*	*			
10		*** 25**	*** 17**	*	*	*		
11		*** 22**	*** 35**		*	*	*** 6**	*** 6**
12		*** 7**	*** 29**	*	*			
13		*** 30**	*** 1**	*	*			
14		*** 13**	*** 10**	*	*			
15		*** 31**	*** 22**		*			
16		*** 6**	*** 20**	*	*			
17		*** 1**	*** 2**	*	*	*		
18		*** 2**	*** 13**	*	*	*		
19		*** 33**	*** 23**		*			
20		*** 3**	*** 26**	*	*			
21		*** 16**	*** 4**	*	*			
22		*** 4**	*** 14**	*	*			
23		*** 5**	*** 33**		*			
24		*** 8**	*** 25**	*	*	*		
25		*** 21**	*** 3**	*	*			
26		*** 26**	*** 11**	*	*			
27		*** 23**	*** 18**		*			
28		*** 35**	*** 24**	*	*	*		
29		*** 34**	*** 6**	*	*	*		
30		*** 18**	*** 12**	*	*			
31		*** 24**	*** 30**		*	*		
32		*** 15**	*** 31**		*			
33		*** 12**	*** 5**	*	*	*	*** 4**	*** 4**
34		*** 32**	*** 15**		*		*** 3**	*** 3**
35		*** 19**	*** 32**		*		*** 2**	*** 2**
36		*** 27**	*** 27**	*	*		*** 1**	*** 1**
classe								
# Atrib.	36	36	36	23	36	12	6	6
% Atrib.		100,00	100,00	63,89	100,00	33,33	16,67	16,67
Média	13,71	13,71	13,71	13,66	13,71	13,55	16,80	16,80
Erro Padrão	0,62	0,62	0,62	0,49	0,62	0,53	0,56	0,56

Tabela B.7: Atributos Selecionados – Satimage

Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
region centroid col		*** 5**	*** 12**		*	*	*** 1**	*** 1**
region centroid row		*** 2**	*** 7**	*	*	*	*** 2**	*** 2**
region pixel count								
short line density 5		*** 15**	*** 15**		*			
short line density 2		*** 14**	*** 16**		*			
vedge mean		*** 7**	*** 14**		*	*		
vegde sd			*** 18**		*			
hedge mean		*** 8**	*** 13**		*	*		
hedge sd		*** 16**	*** 17**		*	*		
intensity mean			*** 4**		*			
rawred mean		*** 3**	*** 5**	*	*	*		
rawblue mean		*** 12**	*** 2**	*	*	*		
rawgreen mean		*** 13**	*** 6**		*			
exred mean		*** 9**	*** 10**		*		*** 3**	*** 3**
exblue mean		*** 10**	*** 9**		*	*		
exgreen mean		*** 6**	*** 8**		*			
value mean		*** 11**	*** 3**		*			
saturatoin mean		*** 4**	*** 11**	*	*		*** 4**	*** 4**
hue mean		*** 1**	*** 1**	*	*	*		
classe								
# Atrib.	19	16	18	5	18	9	4	4
% Atrib.		84,21	94,74	26,32	94,74	47,37	21,05	21,05
Média	3,03	3,46	3,29	3,59	3,29	3,51	6,15	6,15
Erro Padrão	0,35	0,54	0,30	0,25	0,30	0,30	0,35	0,35

Tabela B.8: Atributos Selecionados – Segment

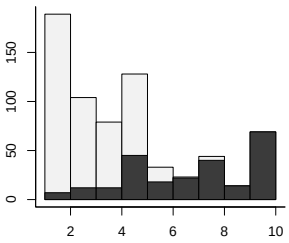
Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
compactness		*** 4**	*** 11**		*	*		
circularity		*** 13**	*** 9**		*	*	*** 4**	*** 4**
distance circularity		*** 9**	*** 6**		*	*	*** 6**	*** 6**
radius ratio		*** 18**	*** 12**	*	*	*		
pr axis aspect ratio		*** 7**	*** 17**	*	*	*		
max length aspect ratio		*** 1**	*** 13**	*	*	*		
scatter ratio		*** 14**	*** 3**	*	*	*		
elongatedness		*** 2**	*** 1**	*	*	*	*** 5**	*** 5**
pr axis rectangularity		*** 8**	*** 5**	*	*	*		
max length rectangularity		*** 10**	*** 7**		*	*		
scaled variance major axis		*** 16**	*** 8**	*	*	*		
scaled variance minor axis		*** 5**	*** 4**	*	*	*		
scaled radius gyration		*** 15**	*** 11**		*	*		
skewness about major axis		*** 6**	*** 15**	*	*	*		
skewness about minor axis		*** 11**	*** 16**	*	*	*	*** 3**	*** 1**
kurtosis about major axis		*** 17**	*** 18**	*	*	*	*** 1**	*** 3**
kurtosis about minor axis		*** 12**	*** 10**		*	*	*** 2**	*** 2**
hollows ratio		*** 3**	*** 2**		*	*		
classe								
# Atrib.	18	18	18	11	18	18	6	6
% Atrib.		100,00	100,00	61,11	100,00	100,00	33,33	33,33
Média	25,90	25,90	25,90	31,68	25,90	25,90	33,92	33,92
Erro Padrão	1,62	1,62	1,62	1,50	1,62	1,62	1,00	1,00

Tabela B.10: Atributos Selecionados – Vehicle

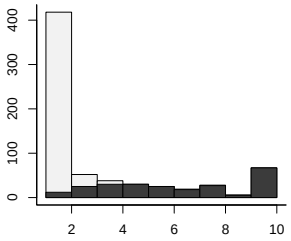
Atributos	Orig.	$\mathcal{C}$4.5	ReliefF	CFS	FCBF	CBF	FDimBF(1)	FDimBF(2)
1		*** 21**	*** 20**					
2		*** 19**	*** 19**		*	*		
3		*** 20**	*** 17**		*	*	*** 2**	*** 2**
4		*** 13**	*** 15**	*	*	*		
5		*** 12**	*** 5**	*	*	*		
6		*** 3**	*** 9**	*	*			
7		*** 7**	*** 6**	*	*	*		
8		*** 15**	*** 13**	*	*	*		
9		*** 6**	*** 11**	*	*	*		
10		*** 8**	*** 3**	*	*	*		
11		*** 1**	*** 1**	*	*	*		
12		*** 4**	*** 2**	*	*	*	*** 1**	*** 1**
13		*** 5**	*** 4**	*	*			
14		*** 11**	*** 8**	*	*			
15		*** 2**	*** 7**	*	*			
16		*** 14**	*** 10**	*	*			
17		*** 10**	*** 12**	*	*	*		
18		*** 9**	*** 14**	*	*		*** 3**	*** 3**
19		*** 17**	**16**		*	*		
20		*** 16**	*** 18**		*			
21		*** 18**	*** 21**					
classe								
# Atrib.	21	21	21	15	19	12	3	3
% Atrib.		100,00	100,00	71,43	90,48	57,14	14,29	14,29
Média	23,00	23,00	23,00	22,38	23,24	24,86	35,16	35,16
Erro Padrão	0,67	0,67	0,67	0,41	0,69	0,88	0,78	0,78

Tabela B.11: Atributos Selecionados – Waveform

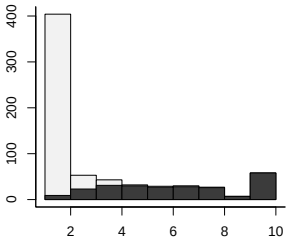
C Distribuição dos Valores dos Atributos dos Conjuntos de Dados



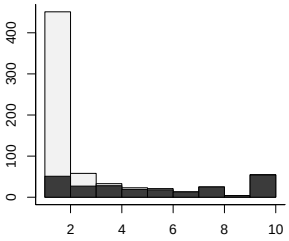
(a) clump thickness



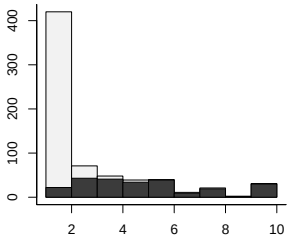
(b) uniformity of cell size



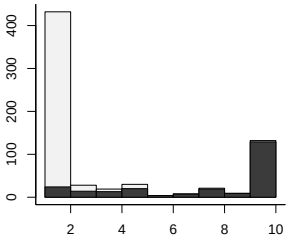
(c) uniformity of cell shape



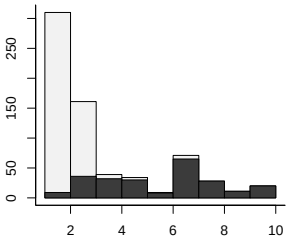
(d) marginal adhesion



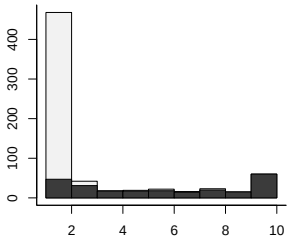
(e) single epithelial cell size



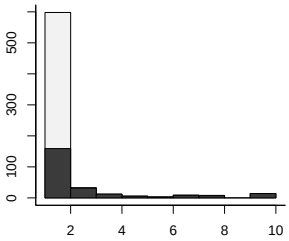
(f) bare nuclei



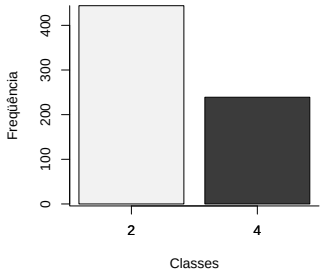
(g) bland chromatin



(h) normal nucleoli

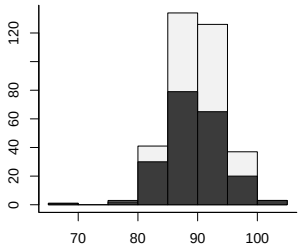


(i) mitoses

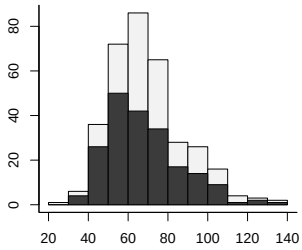


(j) class

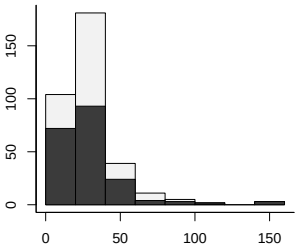
Figura C.23: Distribuições dos valores dos atributos – Breast Cancer



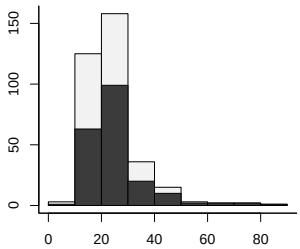
(a) mcv



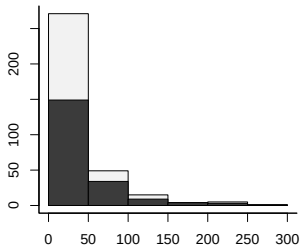
(b) alkphos



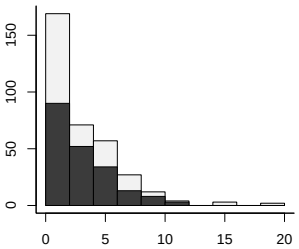
(c) sgpt



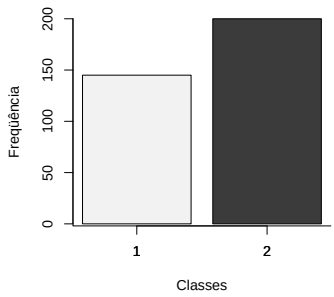
(d) sgot



(e) gammagt

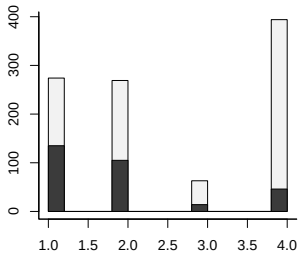


(f) drinks

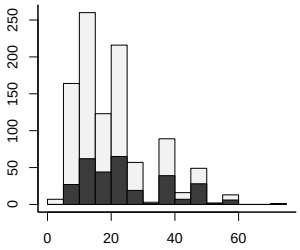


(g) class

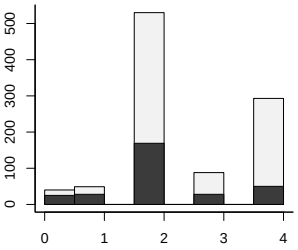
Figura C.24: Distribuições dos valores dos atributos – Bupa



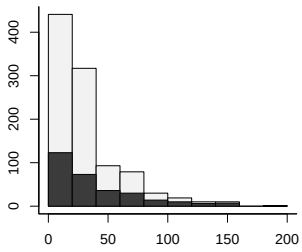
(a) f1



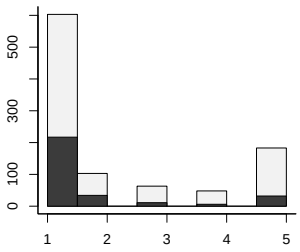
(b) f2



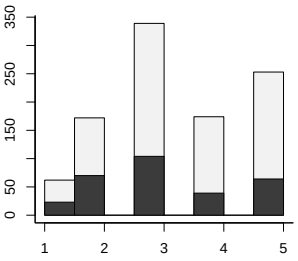
(c) f3



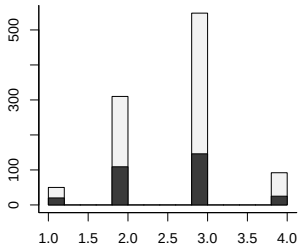
(d) f4



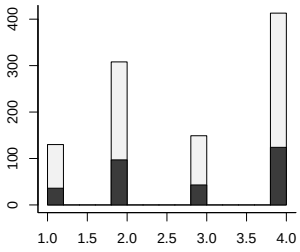
(e) f5



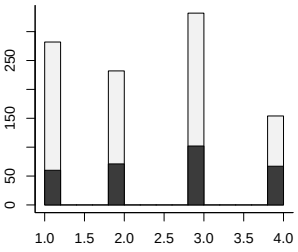
(f) f6



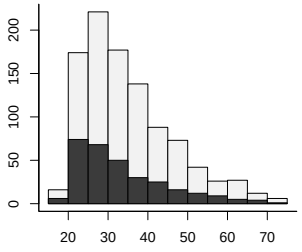
(g) f7



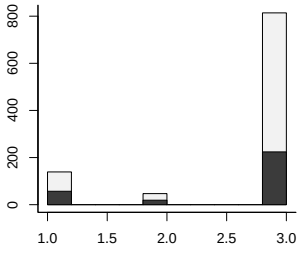
(h) f8



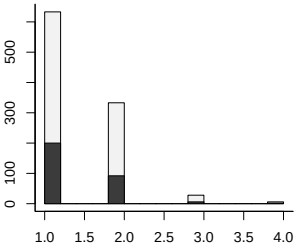
(i) f9



(j) f10

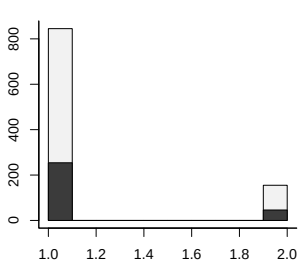


(k) f11

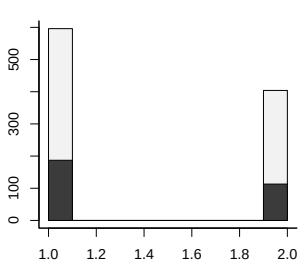


(l) f12

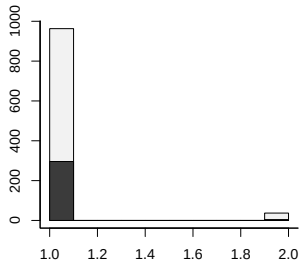
Figura C.25: Distribuições dos valores dos atributos – German – A



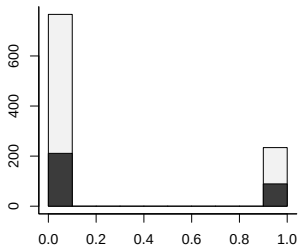
(a) f13



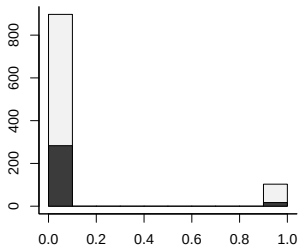
(b) f14



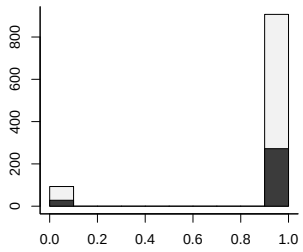
(c) f15



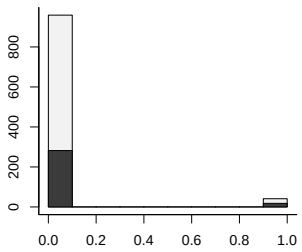
(d) f16



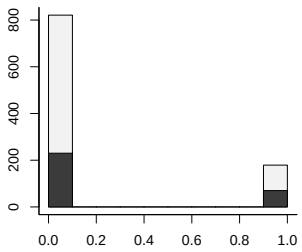
(e) f17



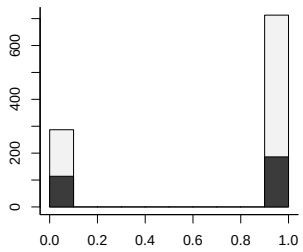
(f) f18



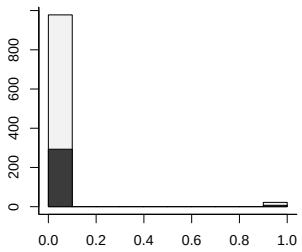
(g) f19



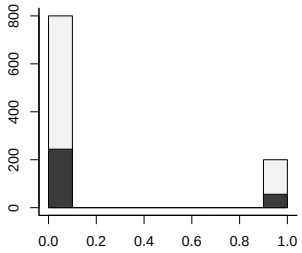
(h) f20



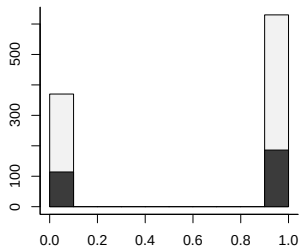
(i) f21



(j) f22

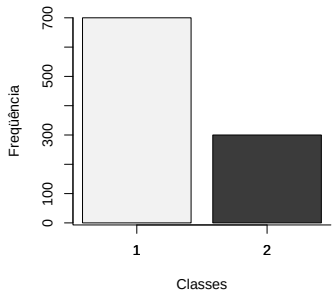


(k) f23



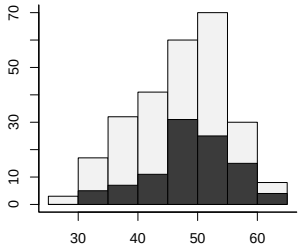
(l) f24

Figura C.26: Distribuições dos valores dos atributos – German – B

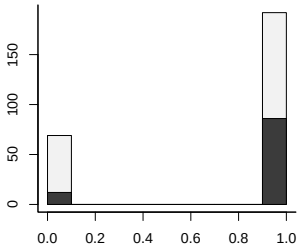


(a) class

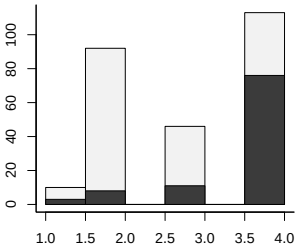
Figura C.27: Distribuições dos valores dos atributos – German – C



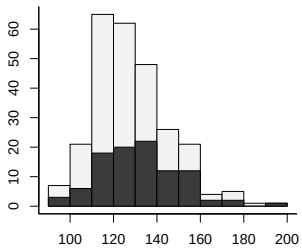
(a) age



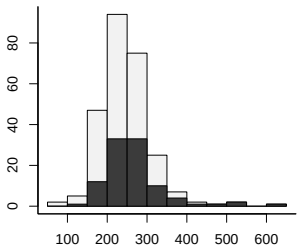
(b) sex



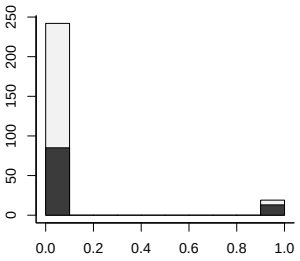
(c) cp



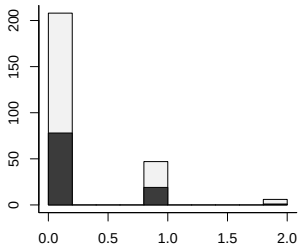
(d) trestbps



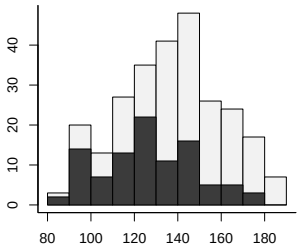
(e) chol



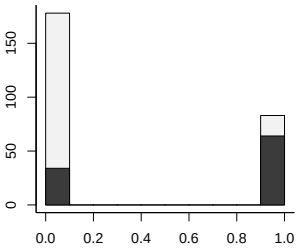
(f) fbs



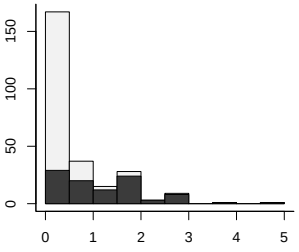
(g) restecg



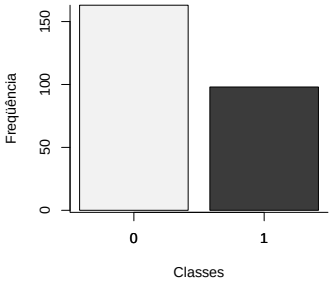
(h) thalach



(i) exang

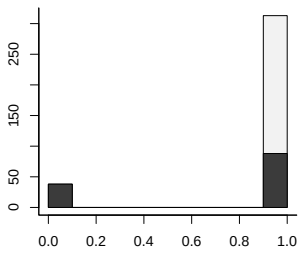


(j) oldpeak

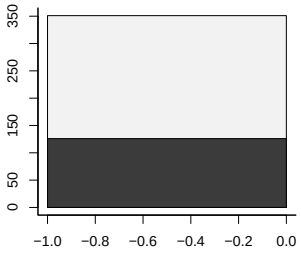


(k) class

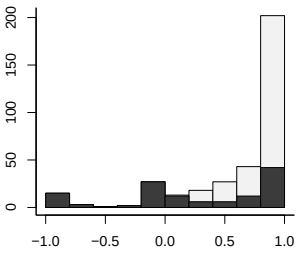
Figura C.28: Distribuições dos valores dos atributos – Hungarian



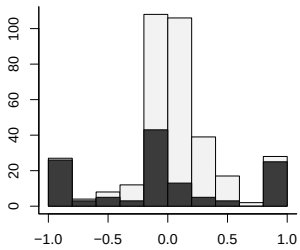
(a) f1



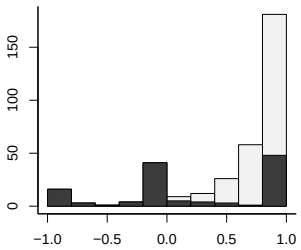
(b) f2



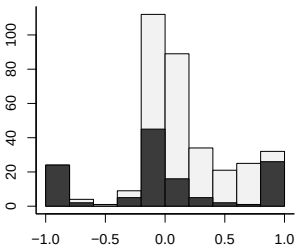
(c) f3



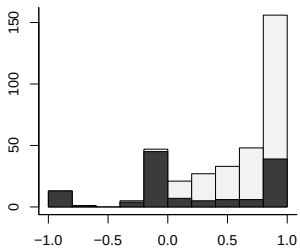
(d) f4



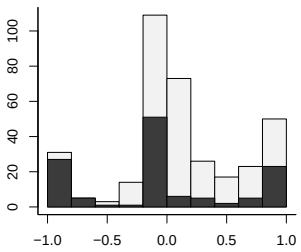
(e) f5



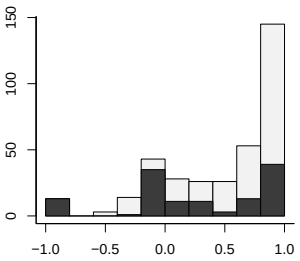
(f) f6



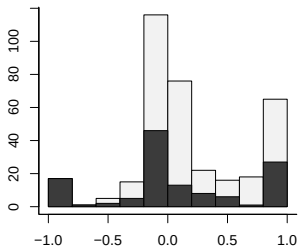
(g) f7



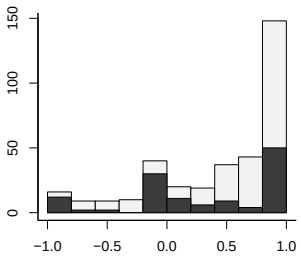
(h) f8



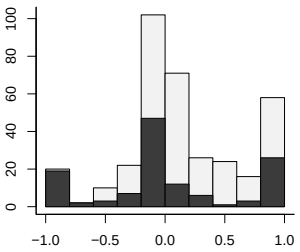
(i) f9



(j) f10

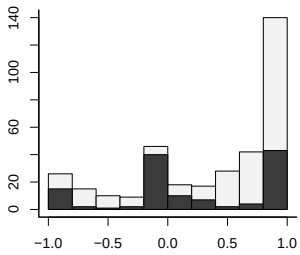


(k) f11

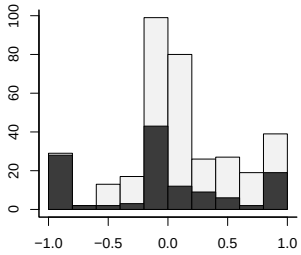


(l) f12

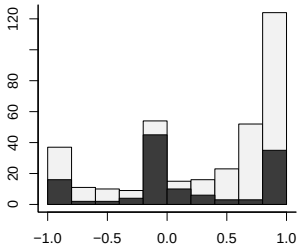
Figura C.29: Distribuições dos valores dos atributos – Ionosphere – A



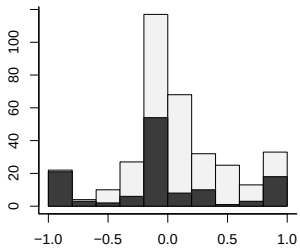
(a) f13



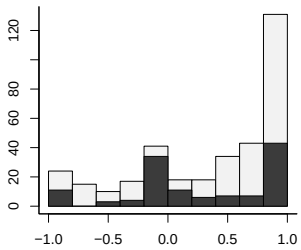
(b) f14



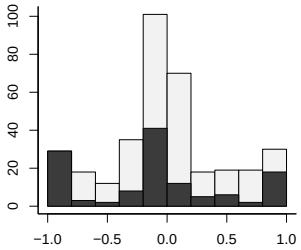
(c) f15



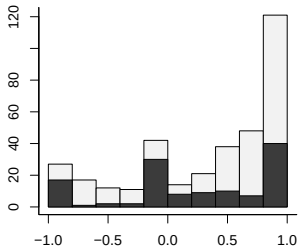
(d) f16



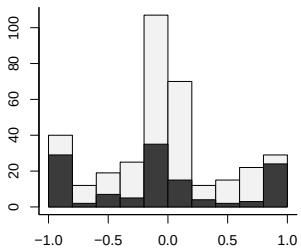
(e) f17



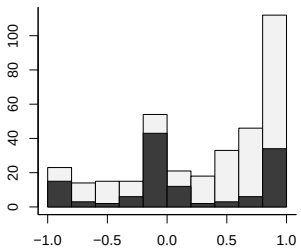
(f) f18



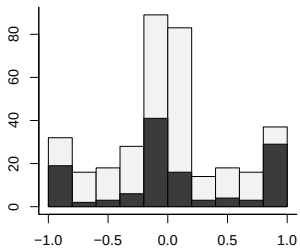
(g) f19



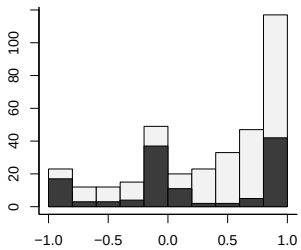
(h) f20



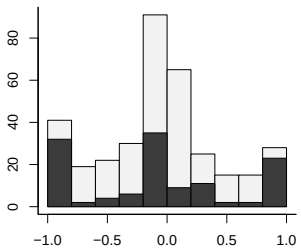
(i) f21



(j) f22

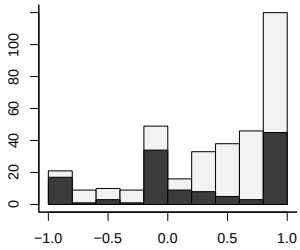


(k) f23

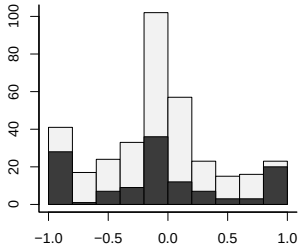


(l) f24

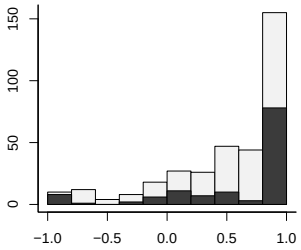
Figura C.30: Distribuições dos valores dos atributos – Ionosphere – B



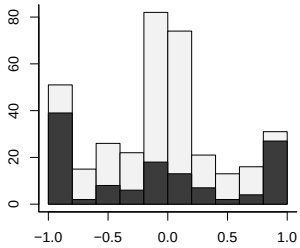
(a) f25



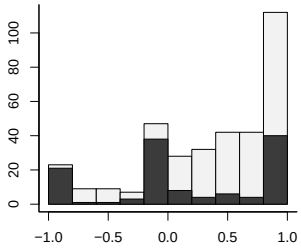
(b) f26



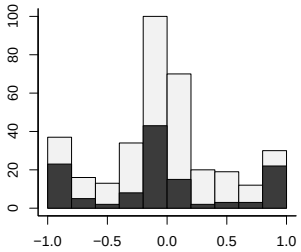
(c) f27



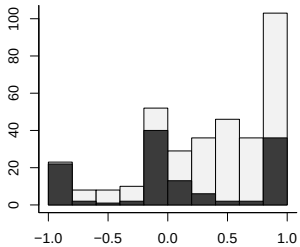
(d) f28



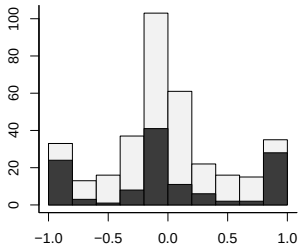
(e) f29



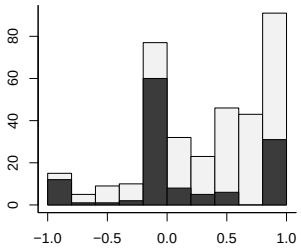
(f) f30



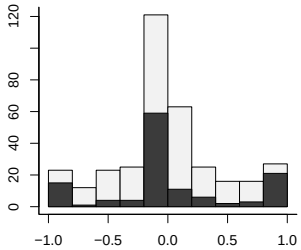
(g) f31



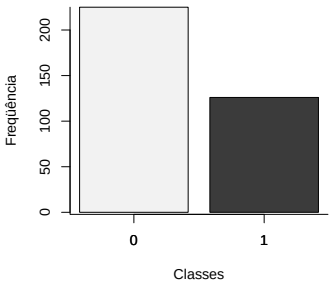
(h) f32



(i) f33

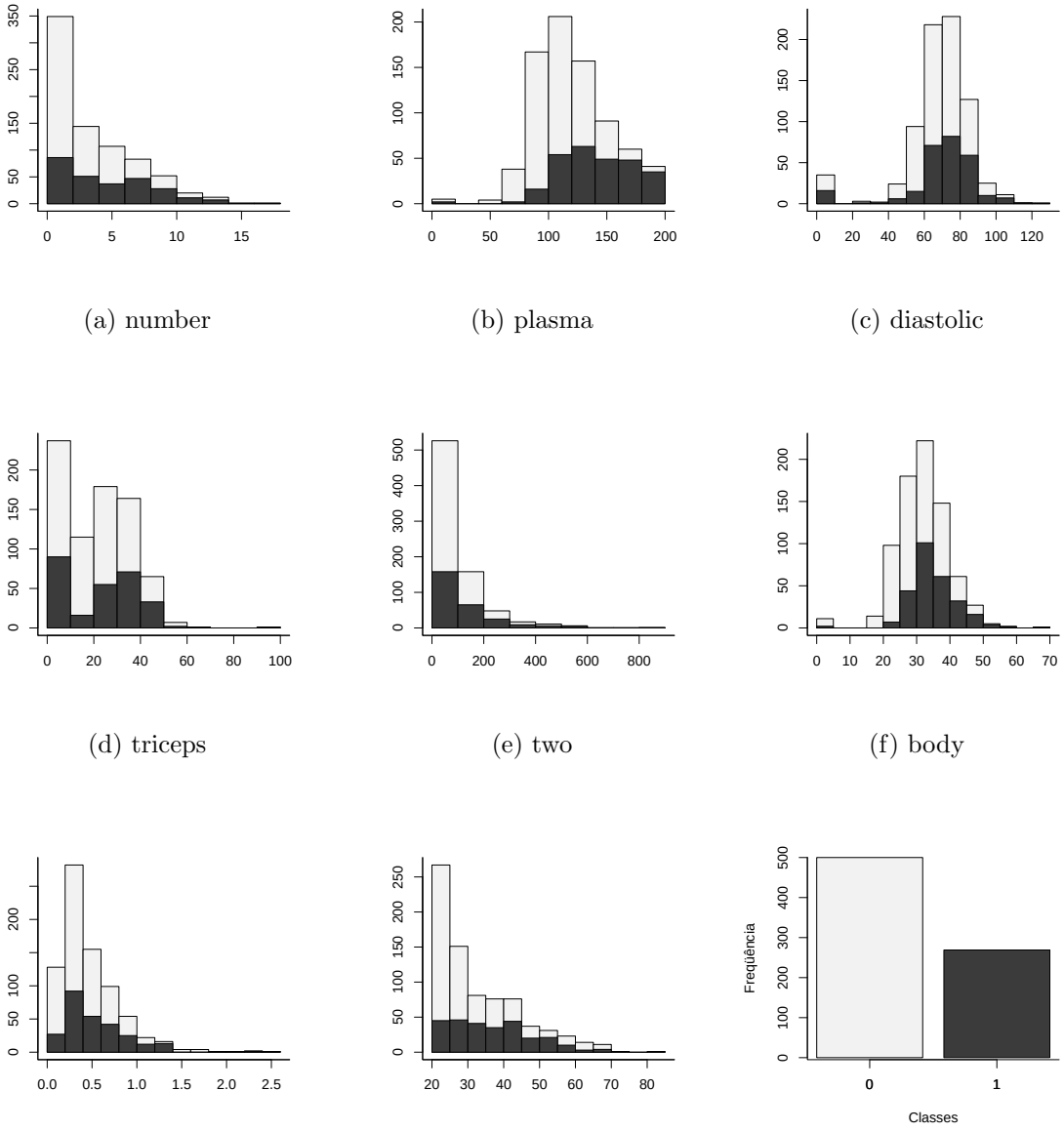


(j) f34



(k) class

Figura C.31: Distribuições dos valores dos atributos – Ionosphere – C

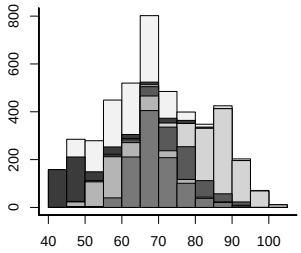


(a) number (b) plasma (c) diastolic

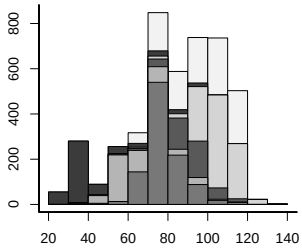
(d) triceps (e) two (f) body

(g) diabetes (h) age (i) class

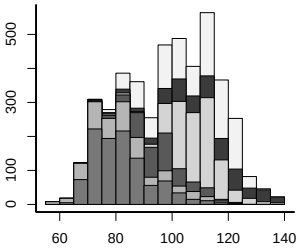
Figura C.32: Distribuições dos valores dos atributos – Pima



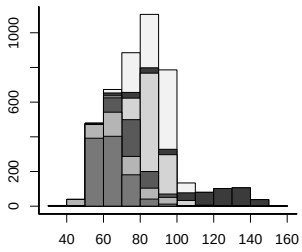
(a) f1



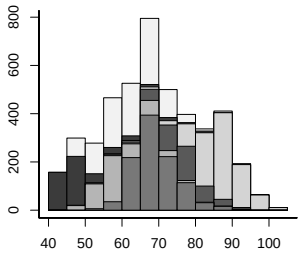
(b) f2



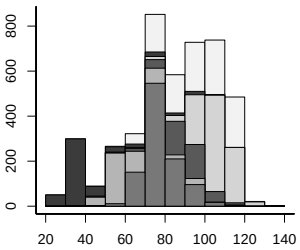
(c) f3



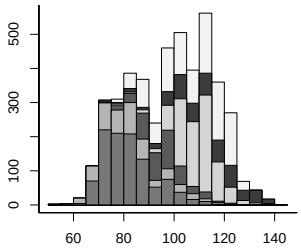
(d) f4



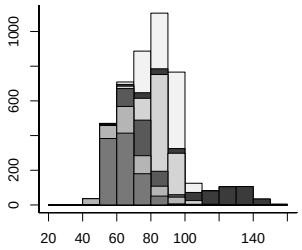
(e) f5



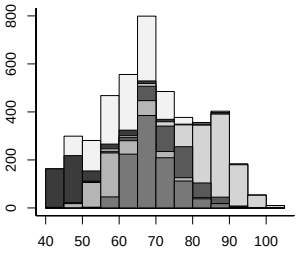
(f) f6



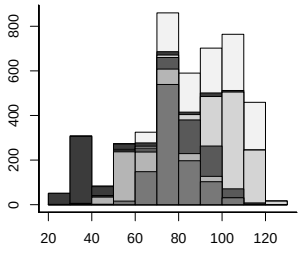
(g) f7



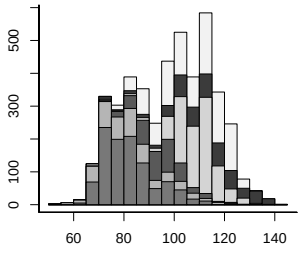
(h) f8



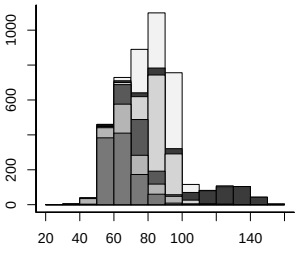
(i) f9



(j) f10

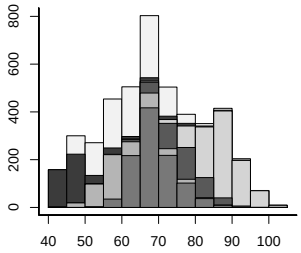


(k) f11

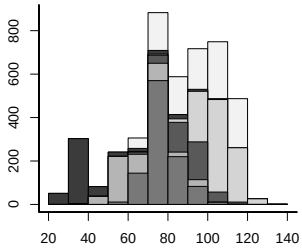


(l) f12

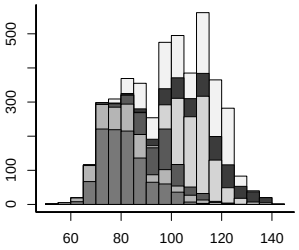
Figura C.33: Distribuições dos valores dos atributos – Satimage – A



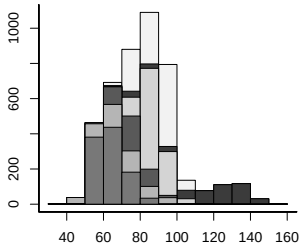
(a) f13



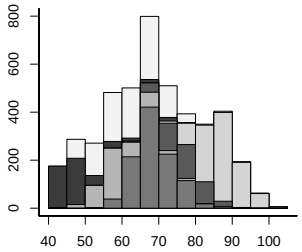
(b) f14



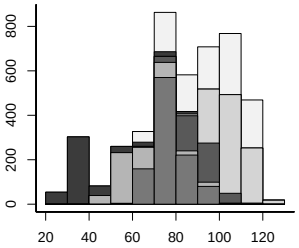
(c) f15



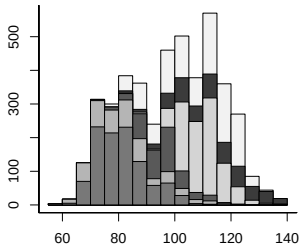
(d) f16



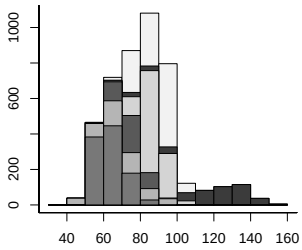
(e) f17



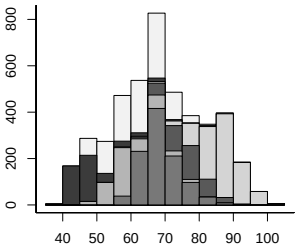
(f) f18



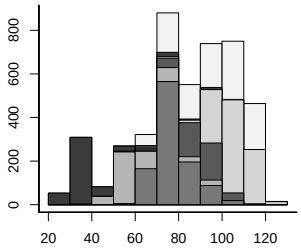
(g) f19



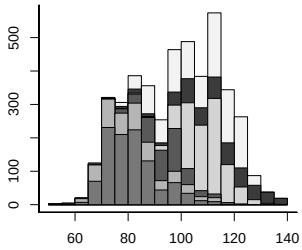
(h) f20



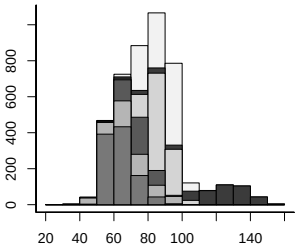
(i) f21



(j) f22

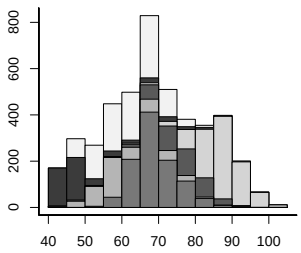


(k) f23

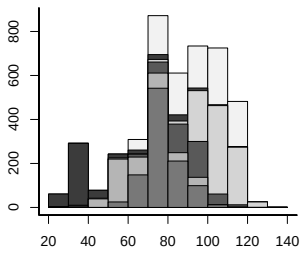


(l) f24

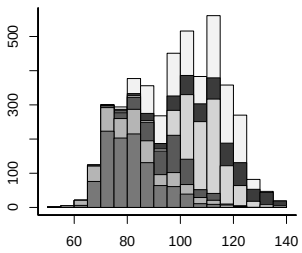
Figura C.34: Distribuições dos valores dos atributos – Satimage – B



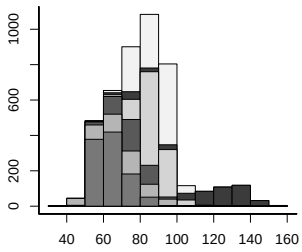
(a) f25



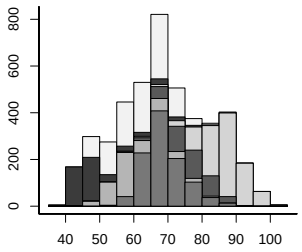
(b) f26



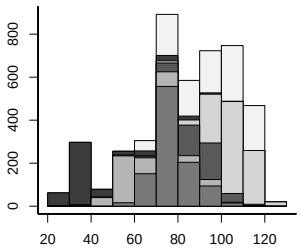
(c) f27



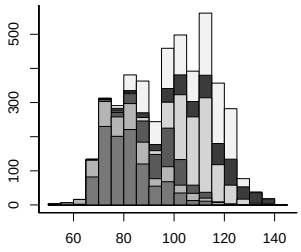
(d) f28



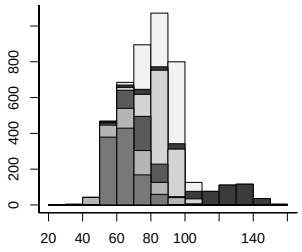
(e) f29



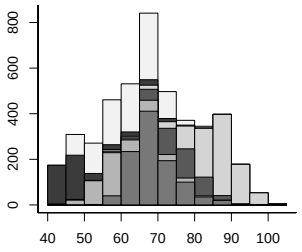
(f) f30



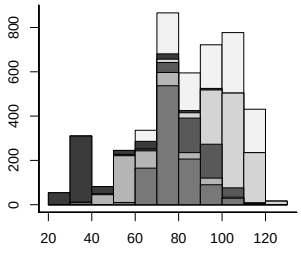
(g) f31



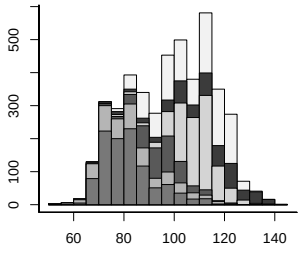
(h) f32



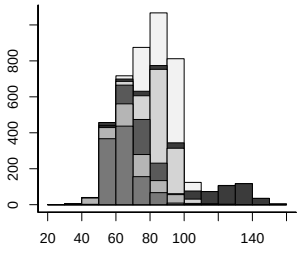
(i) f33



(j) f34

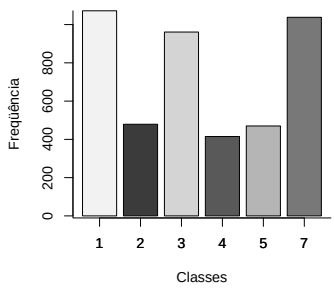


(k) f35



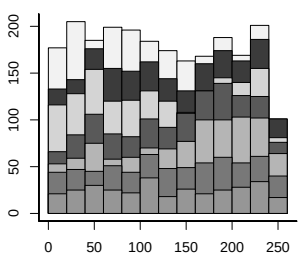
(l) f36

Figura C.35: Distribuições dos valores dos atributos – Satimage – C

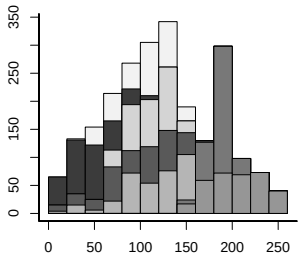


(a) class

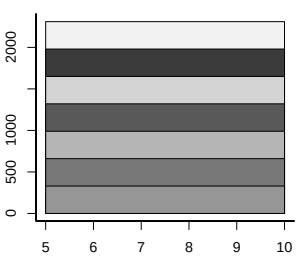
Figura C.36: Distribuições dos valores dos atributos – Satimage – D



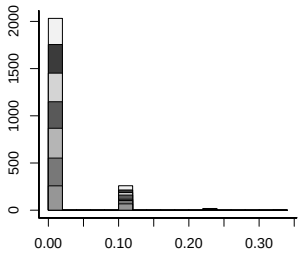
(a) region centroid col



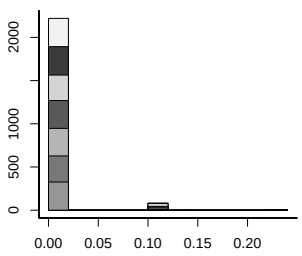
(b) region centroid row



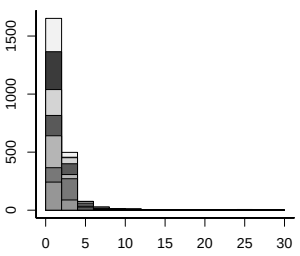
(c) region pixel count



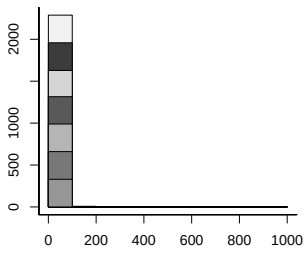
(d) short line density 5



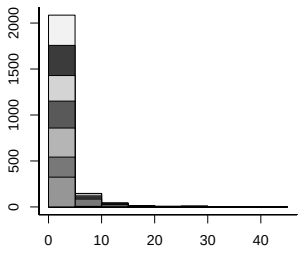
(e) short line density 2



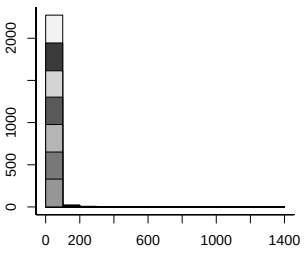
(f) vedge mean



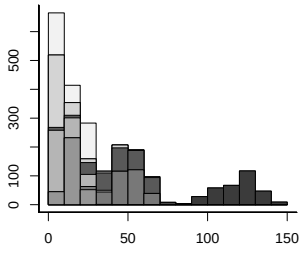
(g) vegde sd



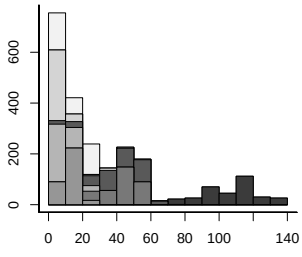
(h) hedge mean



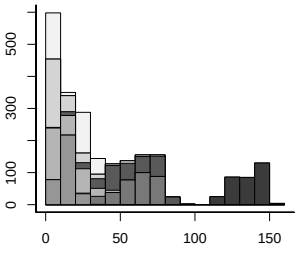
(i) hedge sd



(j) intensity mean

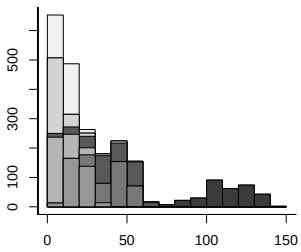


(k) rawred mean

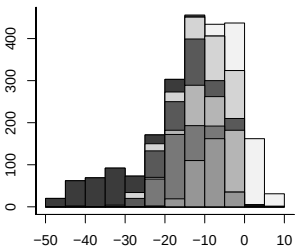


(l) rawblue mean

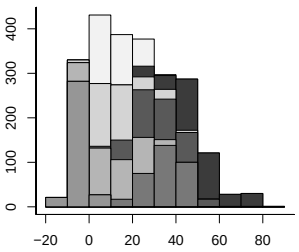
Figura C.37: Distribuições dos valores dos atributos – Segment – A



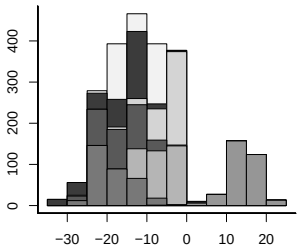
(a) rawgreen mean



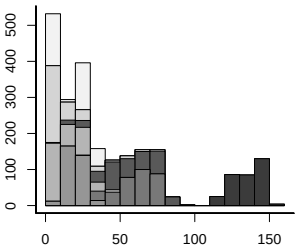
(b) exred mean



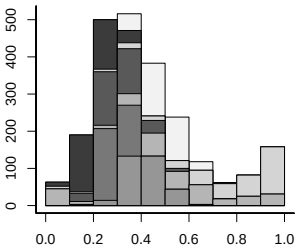
(c) exblue mean



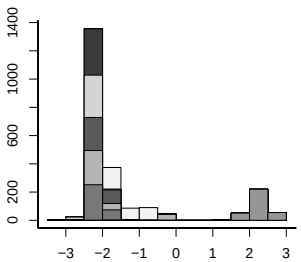
(d) exgreen mean



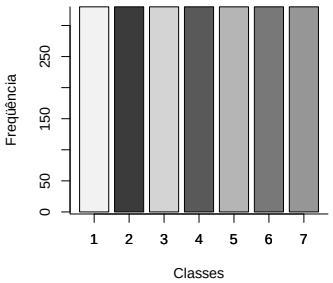
(e) value mean



(f) saturatoin mean

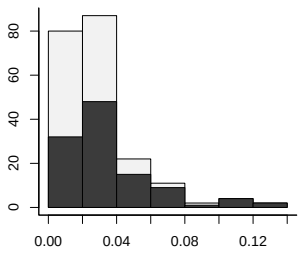


(g) hue mean

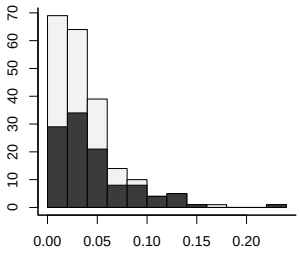


(h) class

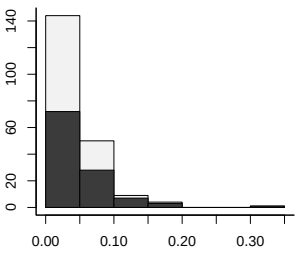
Figura C.38: Distribuições dos valores dos atributos – Segment – B



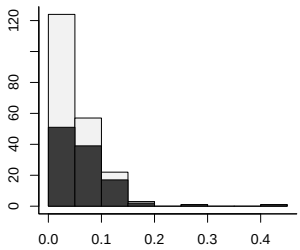
(a) f1



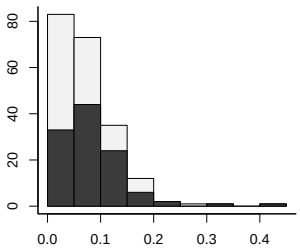
(b) f2



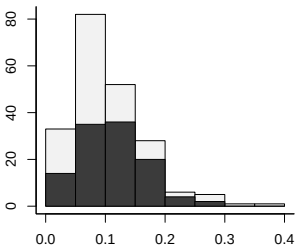
(c) f3



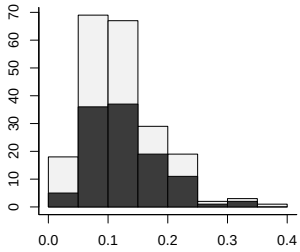
(d) f4



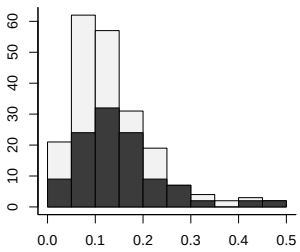
(e) f5



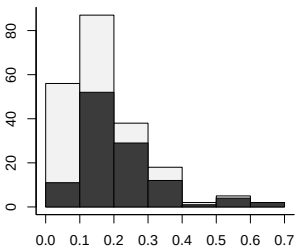
(f) f6



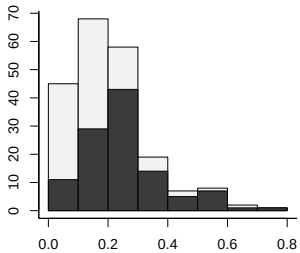
(g) f7



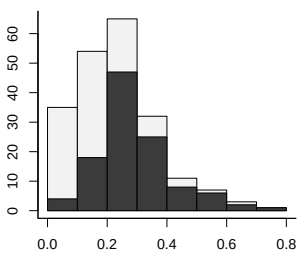
(h) f8



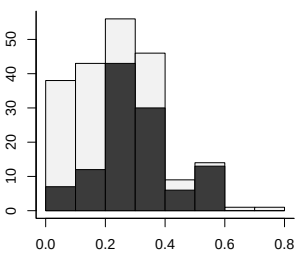
(i) f9



(j) f10

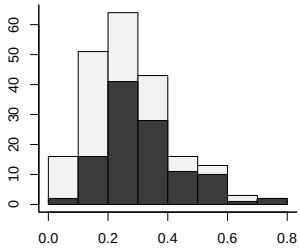


(k) f11

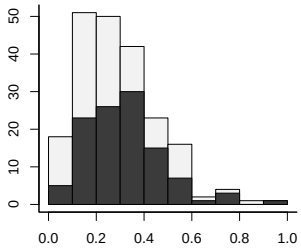


(l) f12

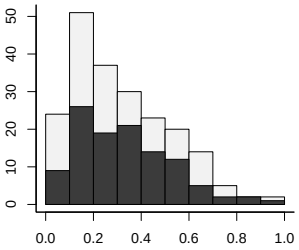
Figura C.39: Distribuições dos valores dos atributos – Sonar – A



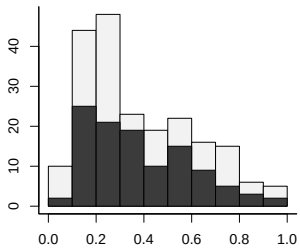
(a) f13



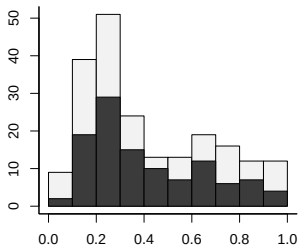
(b) f14



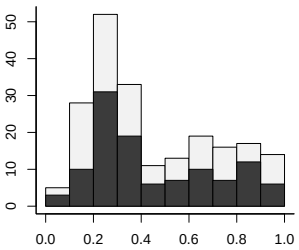
(c) f15



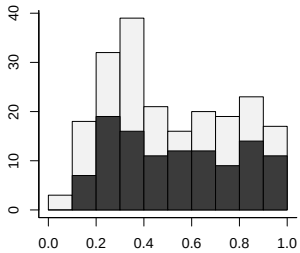
(d) f16



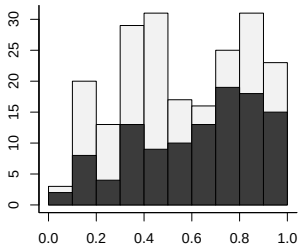
(e) f17



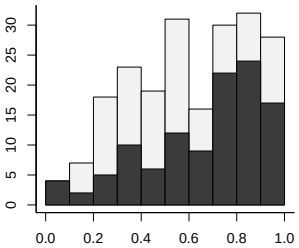
(f) f18



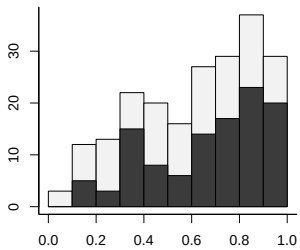
(g) f19



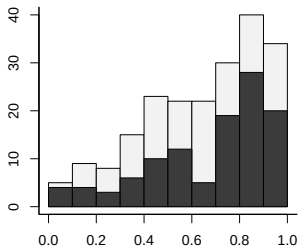
(h) f20



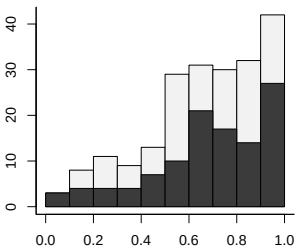
(i) f21



(j) f22

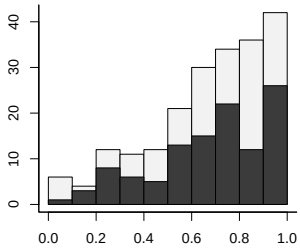


(k) f23

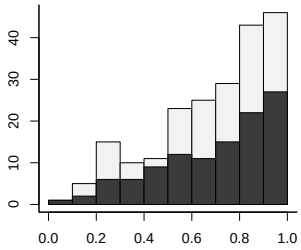


(l) f24

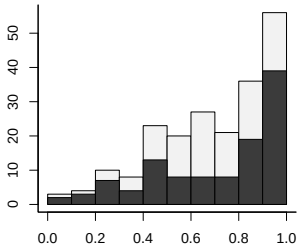
Figura C.40: Distribuições dos valores dos atributos – Sonar – B



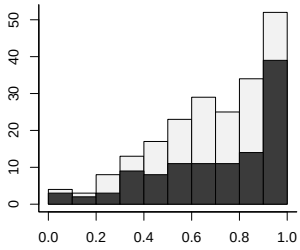
(a) f25



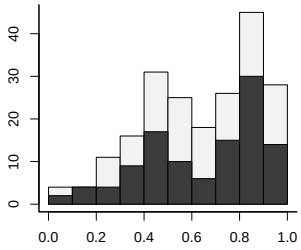
(b) f26



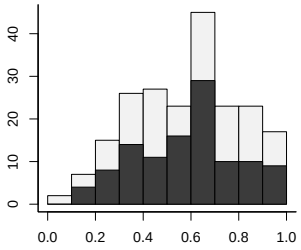
(c) f27



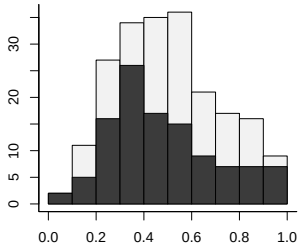
(d) f28



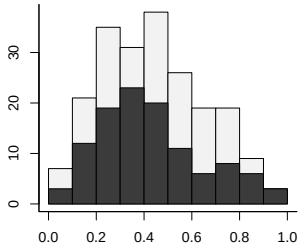
(e) f29



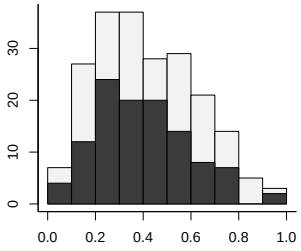
(f) f30



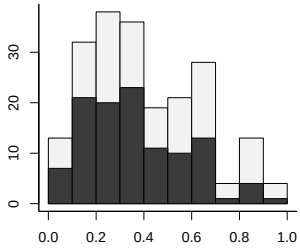
(g) f31



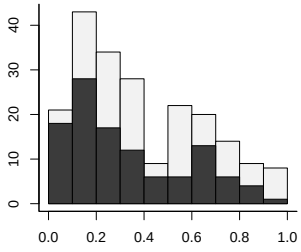
(h) f32



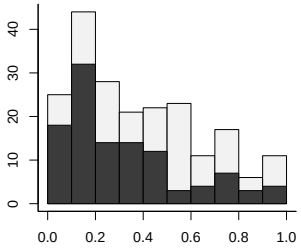
(i) f33



(j) f34

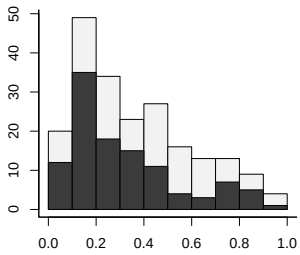


(k) f35

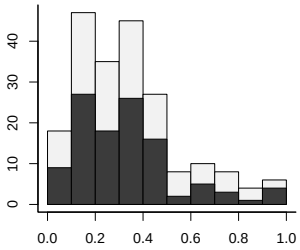


(l) f36

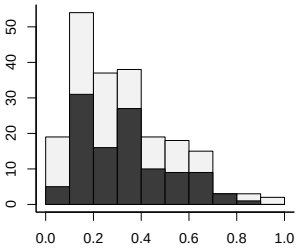
Figura C.41: Distribuições dos valores dos atributos – Sonar – C



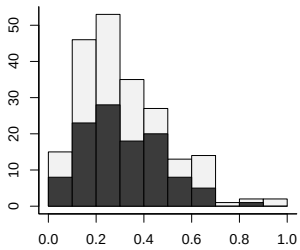
(a) f37



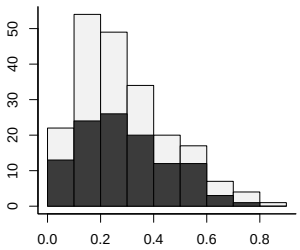
(b) f38



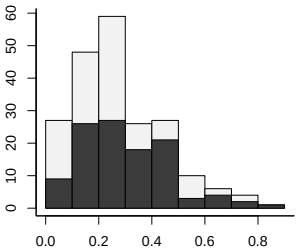
(c) f39



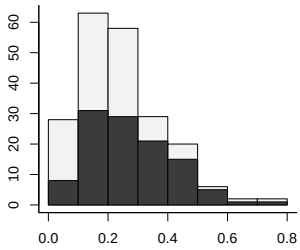
(d) f40



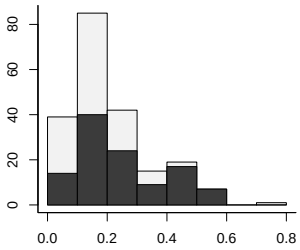
(e) f41



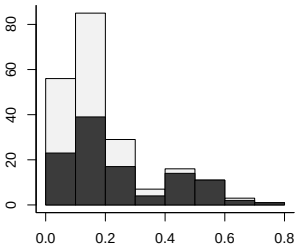
(f) f42



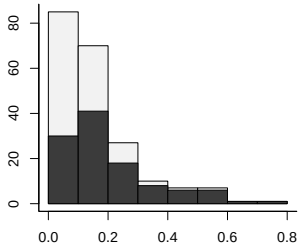
(g) f43



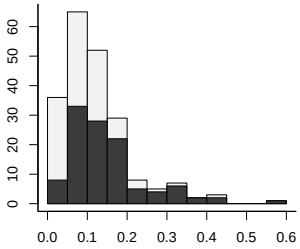
(h) f44



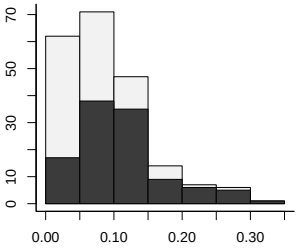
(i) f45



(j) f46

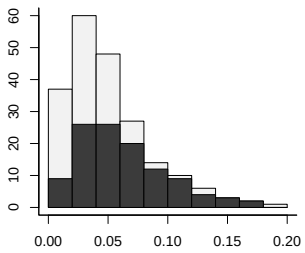


(k) f47

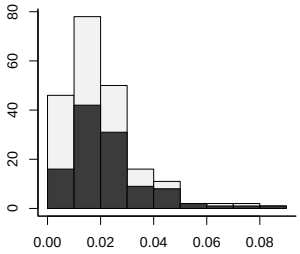


(l) f48

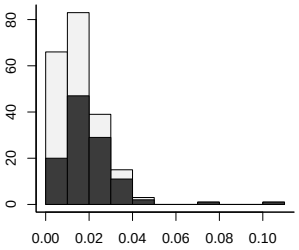
Figura C.42: Distribuições dos valores dos atributos – Sonar – D



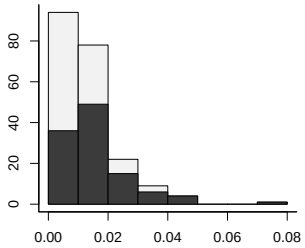
(a) f49



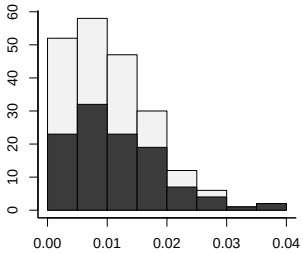
(b) f50



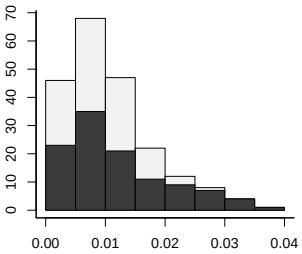
(c) f51



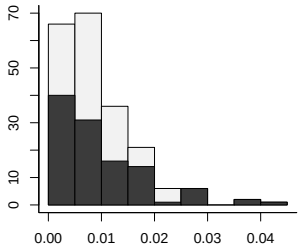
(d) f52



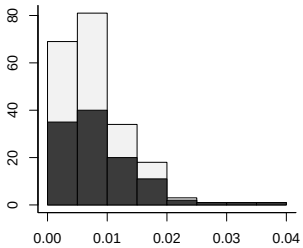
(e) f53



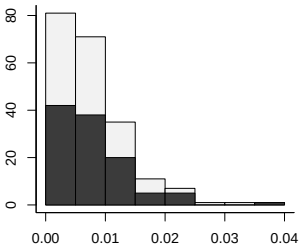
(f) f54



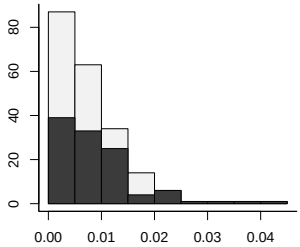
(g) f55



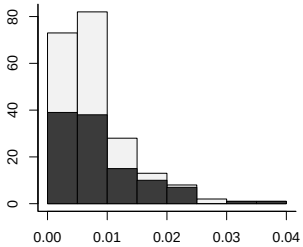
(h) f56



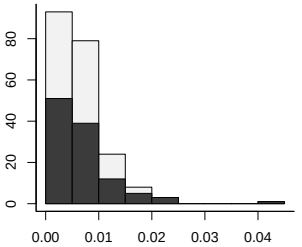
(i) f57



(j) f58

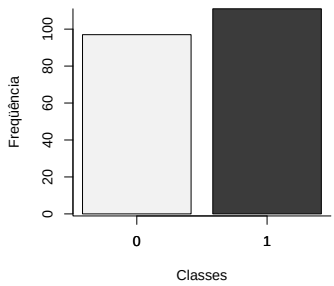


(k) f59



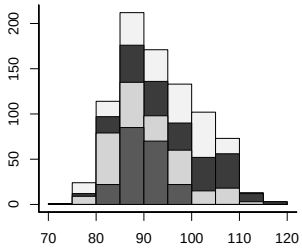
(l) f60

Figura C.43: Distribuições dos valores dos atributos – Sonar – E

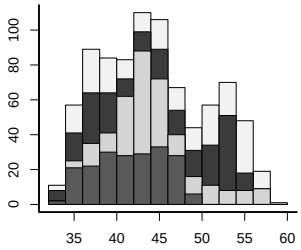


(a) class

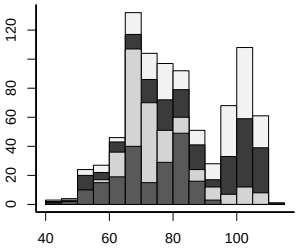
Figura C.44: Distribuições dos valores dos atributos – Sonar – F



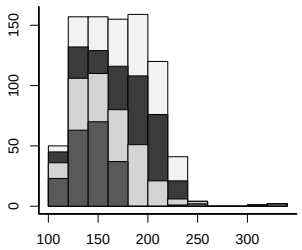
(a) compactness



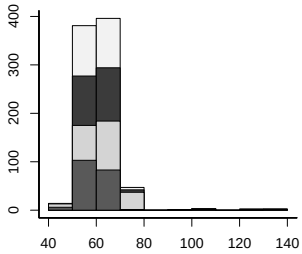
(b) circularity



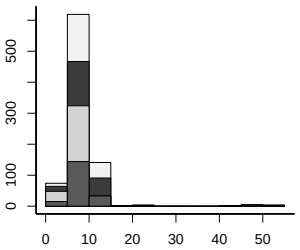
(c) distance circularity



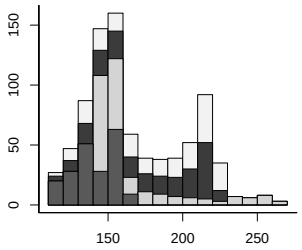
(d) radius ratio



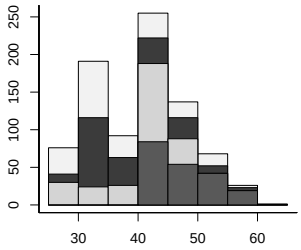
(e) pr axis aspect ratio



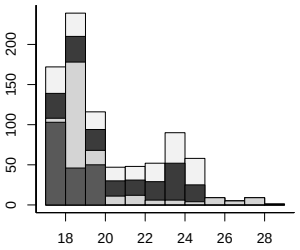
(f) max length aspect ratio



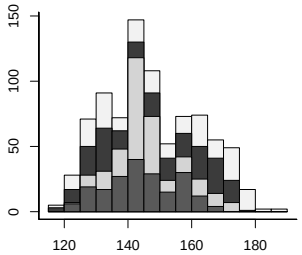
(g) scatter ratio



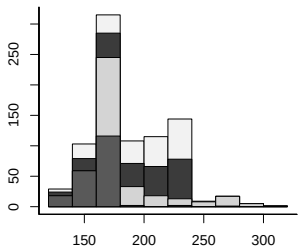
(h) elongatedness



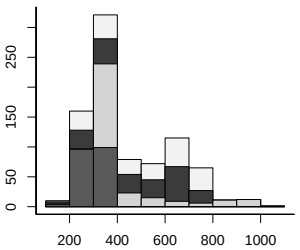
(i) pr axis rectangularity



(j) max length rectangularity

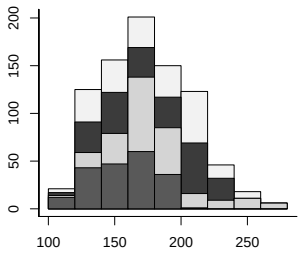


(k) scaled variance major axis

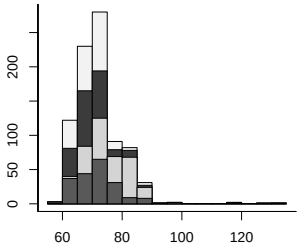


(l) scaled variance minor axis

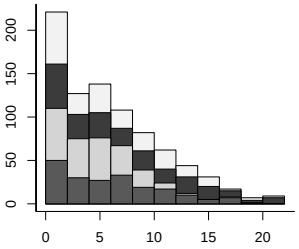
Figura C.45: Distribuições dos valores dos atributos – Vehicle – A



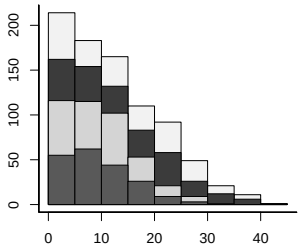
(a) scaled radius gyration



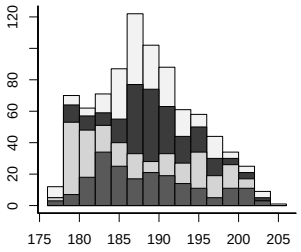
(b) skewness about major axis



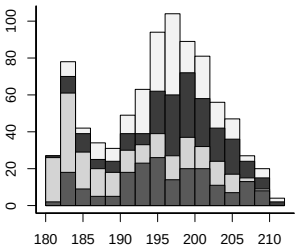
(c) skewness about minor axis



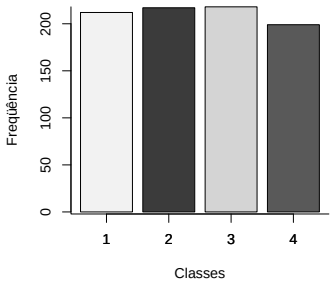
(d) kurtosis about major axis



(e) kurtosis about minor axis

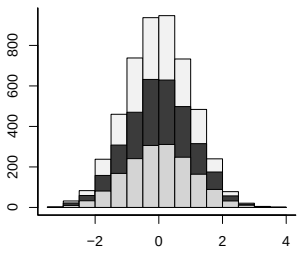


(f) hollows ratio

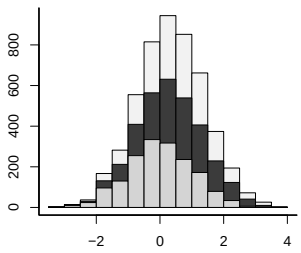


(g) class

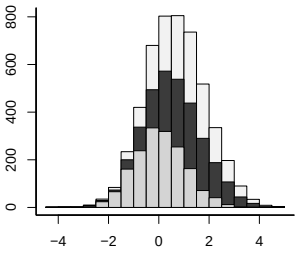
Figura C.46: Distribuições dos valores dos atributos – Vehicle – B



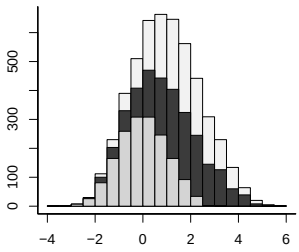
(a) f1



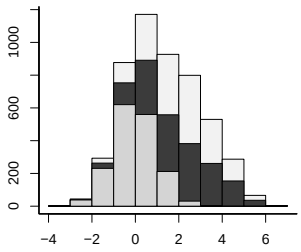
(b) f2



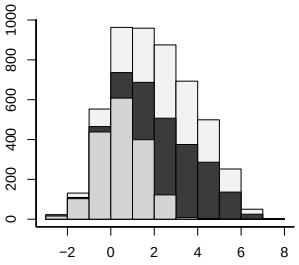
(c) f3



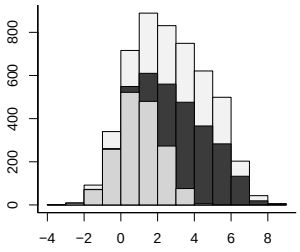
(d) f4



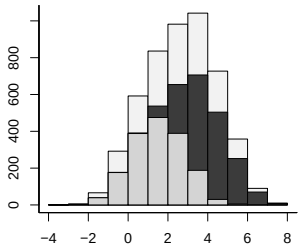
(e) f5



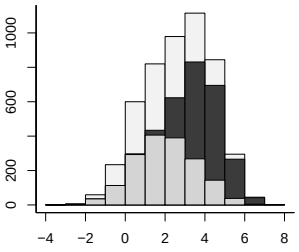
(f) f6



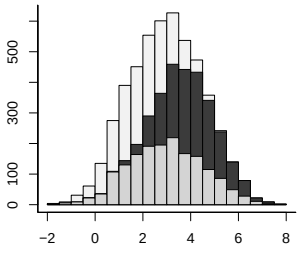
(g) f7



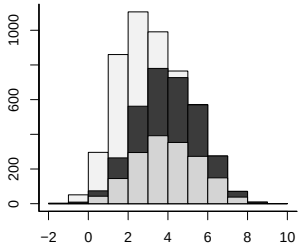
(h) f8



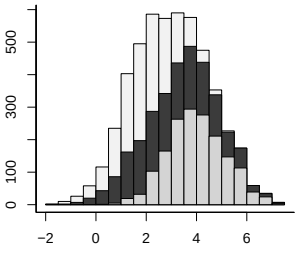
(i) f9



(j) f10

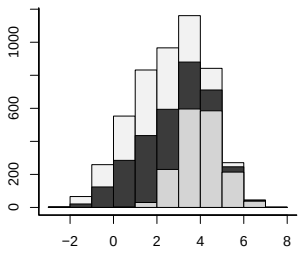


(k) f11

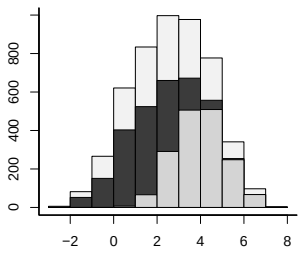


(l) f12

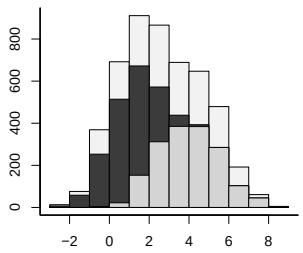
Figura C.47: Distribuições dos valores dos atributos – Waveform – A



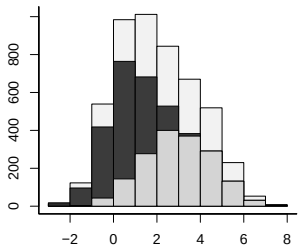
(a) f13



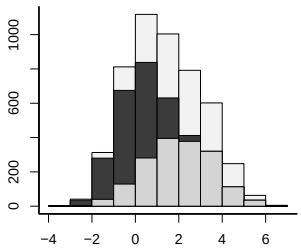
(b) f14



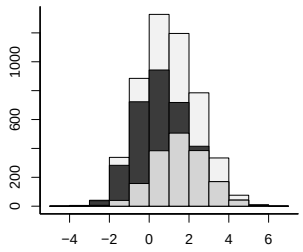
(c) f15



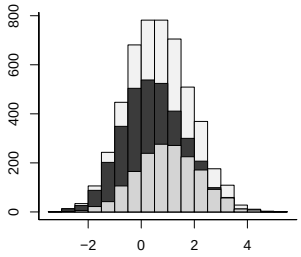
(d) f16



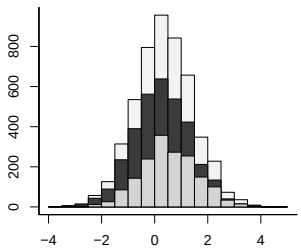
(e) f17



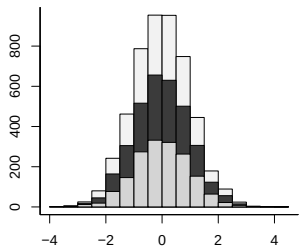
(f) f18



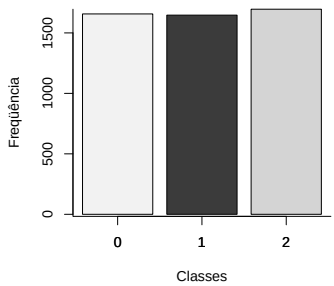
(g) f19



(h) f20



(i) f21



(j) class

Figura C.48: Distribuições dos valores dos atributos – Waveform – B

D Performance dos Algoritmos em Relação à Precisão e a Quantidade de Atributos Selecionados

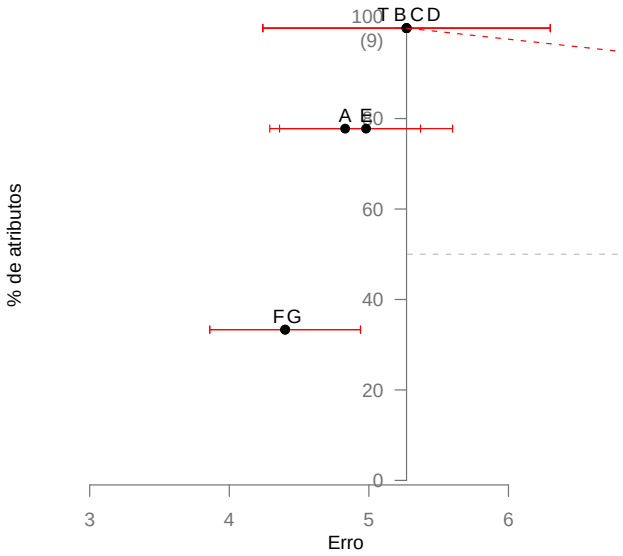


Figura D.49: Gráfico Percentagem × Erro - Breast Cancer

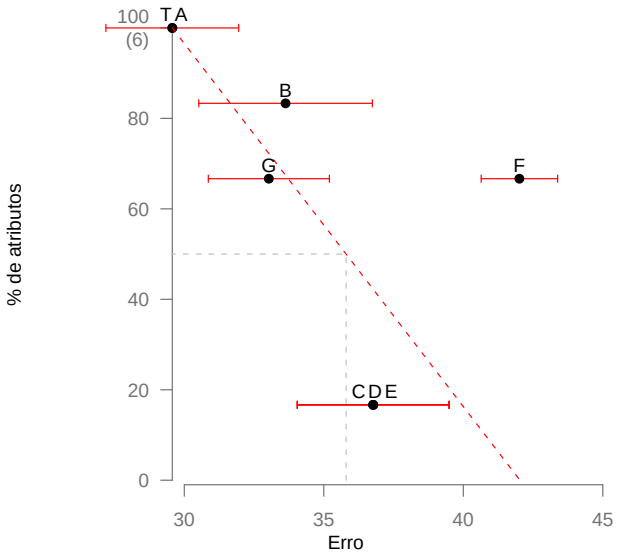


Figura D.50: Gráfico percentagem × erro - Bupa

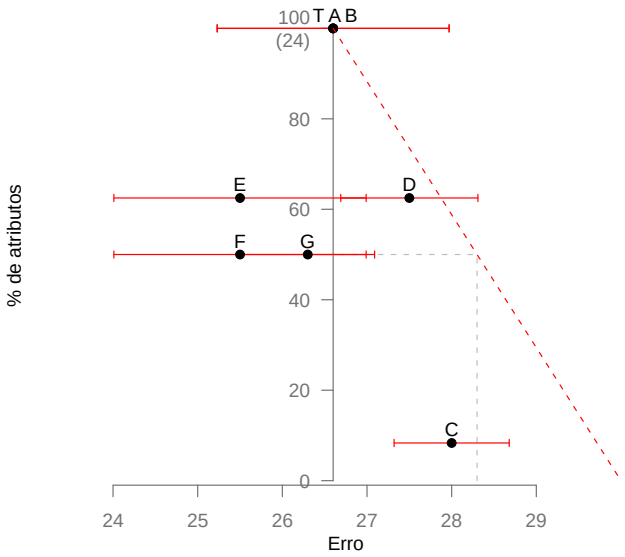


Figura D.51: Gráfico percentagem × erro - German

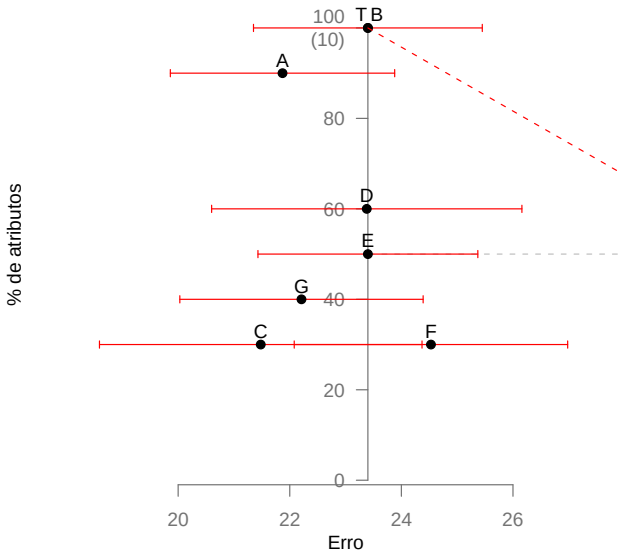


Figura D.52: Gráfico percentagem × erro - Hungarian

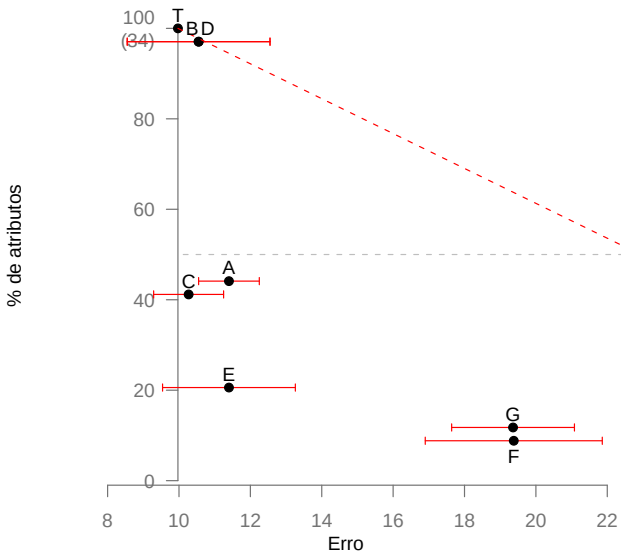


Figura D.53: Gráfico percentagem × erro - Ionosphere

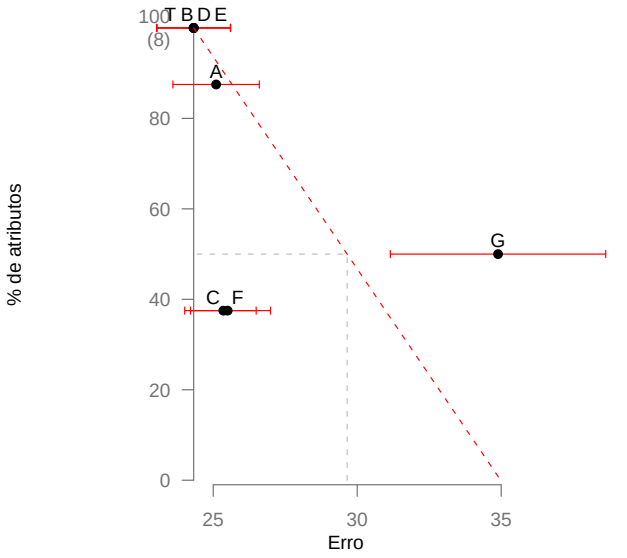


Figura D.54: Gráfico percentagem × erro - Pima

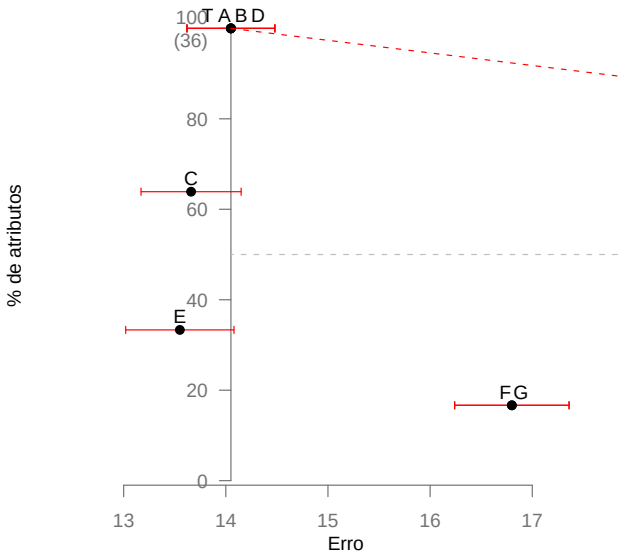


Figura D.55: Gráfico percentagem × erro - Satimage

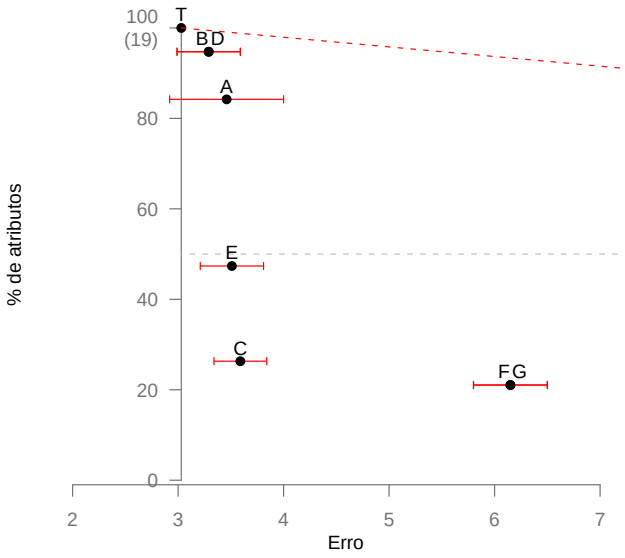


Figura D.56: Gráfico percentagem × erro - Segment

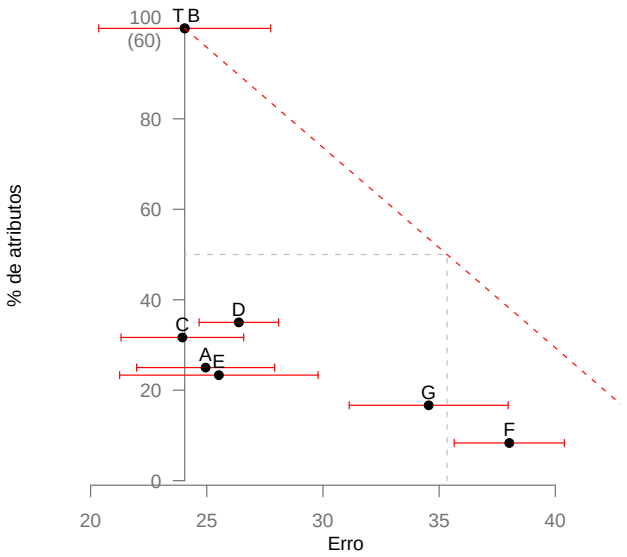


Figura D.57: Gráfico percentagem × erro - Sonar

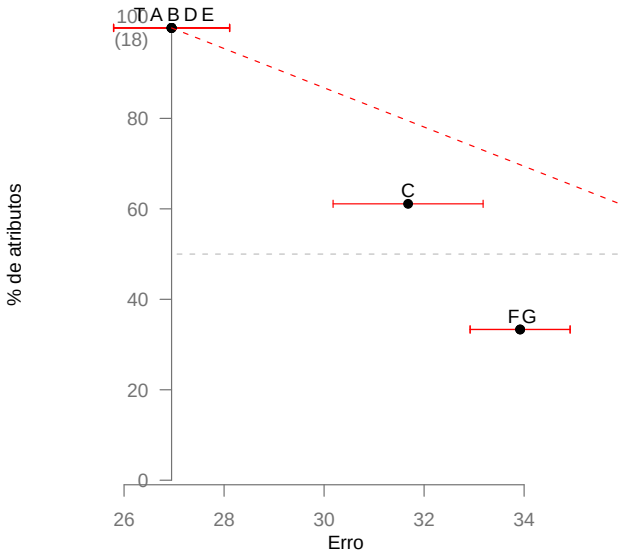


Figura D.58: Gráfico percentagem × erro - Vehicle

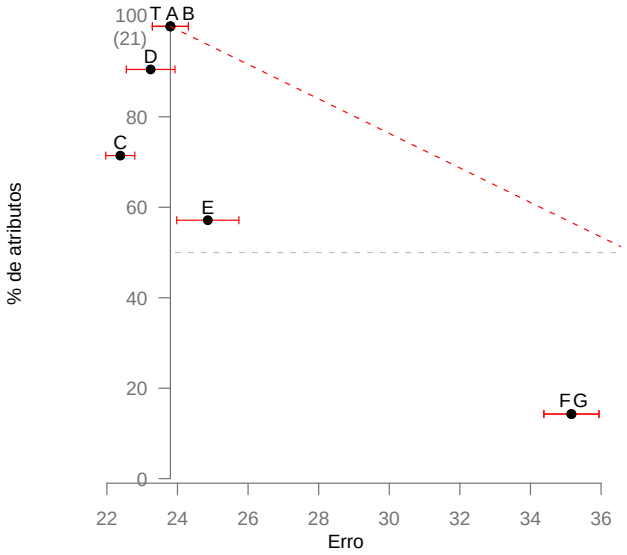


Figura D.59: Gráfico percentagem × erro - Waveform

E Regras Induzidas com o Conjunto de Dados Meta1

```
See5 [Release 1.19]     Fri Sep 09 20:26:22 2005
-------------------

    Options:
    Rule-based classifiers

Read 132 cases (11 attributes) from dados_df.data

Rules:

Rule 1: (18, lift 4.2)
    pontos <= 6
    proporcao > 10.32
    proporcao <= 41.67
    ->  class excelente  [0.950]

Rule 2: (36/12, lift 2.9)
    formato_orig = 1
    ->  class excelente  [0.658]

Rule 3: (36, lift 1.6)
    proporcao > 96.13
    ->  class muito_bom  [0.974]

Rule 4: (12, lift 1.6)
    formato_orig = 2
    ->  class muito_bom  [0.929]

Rule 5: (84/30, lift 1.1)
    formato_orig = 3
    ->  class muito_bom  [0.640]

Rule 6: (6, lift 19.2)
    pontos > 6
    proporcao > 41.67
    formato_orig = 3
    ->  class bom  [0.875]

Rule 7: (6, lift 19.2)
    fractal_3 = bom
    proporcao <= 10.32
    formato_orig = 3
    ->  class regular  [0.875]

Rule 8: (6, lift 9.6)
    fractal_3 = bom
    pontos <= 6
    proporcao > 41.67
```

```
    ->  class ruim  [0.875]

Default class: muito_bom


Evaluation on training data (132 cases):

            Rules
      ----------------
        No      Errors

         8    6( 4.5%)   <<




       (a)   (b)   (c)   (d)   (e)    <-classified as
      ----  ----  ----  ----  ----
        30                             (a): class excelente
              78                       (b): class muito_bom
                    6                  (c): class bom
                          6            (d): class regular
               6                 6     (e): class ruim

Time: 0.1 secs
```

F Regras Induzidas com o Conjunto de Dados Meta2

```
See5 [Release 1.19]     Fri Sep 09 15:59:04 2005
-------------------

    Options:
    Rule-based classifiers

Read 132 cases (11 attributes) from dados_df.data

Rules:

Rule 1: (24, lift 4.2)
    diferenca = nao
    formato_orig = 1
    ->  class excelente  [0.962]

Rule 2: (6, lift 3.8)
    pontos <= 6
    proporcao > 10.32
    proporcao <= 26.1
    ->  class excelente  [0.875]

Rule 3: (24, lift 1.6)
    fractal_3 = medio
    pontos > 5
    ->  class muito_bom  [0.962]

Rule 4: (24, lift 1.6)
    fractal_3 = ruim
    formato_orig = 3
    ->  class muito_bom  [0.962]

Rule 5: (18, lift 1.6)
    abordagem2 = FDimBF_C45
    pontos <= 5
    proporcao > 26.1
    formato_orig = 3
    ->  class muito_bom  [0.950]

Rule 6: (12, lift 1.6)
    formato_orig = 2
    ->  class muito_bom  [0.929]

Rule 7: (6, lift 1.5)
    pontos > 6
    proporcao <= 26.1
    ->  class muito_bom  [0.875]

Rule 8: (6, lift 1.5)
```

```
    pontos <= 2
    ->  class muito_bom  [0.875]

Rule 9: (6, lift 19.2)
    pontos > 6
    proporcao > 26.1
    formato_orig = 3
    ->  class bom  [0.875]

Rule 10: (6, lift 19.2)
    pontos > 2
    proporcao <= 10.32
    formato_orig = 3
    ->  class regular  [0.875]

Rule 11: (6, lift 9.6)
    fractal_3 = bom
    pontos <= 6
    proporcao > 26.1
    ->  class ruim  [0.875]

Rule 12: (6, lift 9.6)
    abordagem2 = FDimBF_ReliefF
    fractal_3 = medio
    pontos <= 5
    proporcao > 26.1
    formato_orig = 3
    ->  class ruim  [0.875]

Default class: muito_bom


Evaluation on training data (132 cases):

            Rules
      ----------------
        No      Errors

        12    0( 0.0%)   <<


       (a)   (b)   (c)   (d)   (e)    <-classified as
      ----  ----  ----  ----  ----
        30                             (a): class excelente
              78                       (b): class muito_bom
                     6                 (c): class bom
                           6           (d): class regular
                                12     (e): class ruim

Time: 0.0 secs
```