UNIVERSIDADE FEDERAL DO PARANÁ

MONICA BELTRAMI

MÉTODO *GRID-QUADTREE* PARA A SELEÇÃO DE PARÂMETROS DO ALGORITMO *SUPPORT VECTOR CLASSIFICATION* (SVC)

CURITIBA

2016

MONICA BELTRAMI

MÉTODO *GRID-QUADTREE* PARA A SELEÇÃO DE PARÂMETROS DO ALGORITMO *SUPPORT VECTOR CLASSIFICATION* (SVC)

Tese apresentada ao Curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, Setores de Tecnologia e Ciência Exatas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciências.

Orientador: Prof. Dr. Arinei Carlos Lindbeck da Silva.

CURITIBA

2016

Beltrami, Monica
Método grid-quadtree para a seleção de parâmetros do algoritmo support vector classification (SVC) / Monica Beltrami. – Curitiba, 2016.
193 f. : il., tabs.

Tese (doutorado) – Universidade Federal do Paraná, Setores de Tecnologia e Ciências Exatas, Programa de Pós-Graduação em Métodos numéricos em Engenharia.
Orientador: Arinei Carlos Lindbeck da Silva
Bibliografia: p.143-149

1. Algoritmos. 2. Sistemas de reconhecimento de padrões Estimativa de parâmetros. I. Silva, Arinei Carlos Lindbeck. II. Título.

CDD 610.28

Ministério da Educação
Universidade Federal do Paraná
Setor de Tecnologia / Setor de Ciências Exatas
Departamento de Construção Civil / Departamento de Matemática/ Departamento de Engenharia de Produção.
Programa de Pós-Graduação em Métodos Numéricos em Engenharia – PPGMNE/UFPR.

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de MÔNICA BELTRAMI, intitulada: "MÉTODO GRID-QUADTREE PARA A SELEÇÃO DE PARÂMETROS DO ALGORÍTMO SUPPORT VECTOR CLASSIFICATION (SVC)", após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 01 de junho de 2016.

ARINEI CARLOS LINDBECK DA SILVA (UFPR)
(Presidente da Banca Examinadora)

CASSIUS TADEU SCARPIN (UFPR)

DEISE MARIA BERTHOLDI COSTA (UFPR)

MARCO ANTÔNIO LUERSEN (UTFPR)

ALEXANDRE RASI AOKI (UFPR)

LUCAS GARCIA PEDROSO (UFPR)

Curitiba, 01 de junho de 2016

Aos meus pais, por todo
amor e incentivo.

AGRADECIMENTOS

A Deus, pela vida, bênçãos e proteção.

Aos meus pais, Renato Beltrami e Sandra Elizabeth Hromada Beltrami, pelo amor, incentivo e apoio em todos os momentos da minha vida.

Aos meus irmãos, Katia Beltrami Koppe e Enzo Beltrami, pelo apoio e amizade.

Ao meu orientador, Prof. Dr. Arinei Carlos Lindbeck da Silva, pelos ensinamentos, orientações, amizade e principalmente pela paciência.

Aos professores do PPGMNE, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, pelos conhecimentos transmitidos e pela oportunidade.

Ao meu namorado Gustavo Valentim Loch, pelo amor, amizade, apoio e contribuições.

Aos meus amigos e colegas de curso, Crisiane Rezende Vilela de Oliveira e Paulo Amaro Velloso Henriques dos Santos, pela amizade, apoio e momentos de estudos em grupo.

Ao Instituto Federal do Paraná – IFPR, pela licença capacitação, que possibilitou dedicação integral aos estudos.

À CAPES, pelo auxílio financeiro.

Ao GTAO, Grupo de Tecnologia Aplicada à Otimização, pelos recursos disponibilizados.

A todas as outras pessoas, que direta ou indiretamente contribuíram para a realização deste trabalho.

RESUMO

O algoritmo *Support Vector Classification* (SVC) é uma técnica de reconhecimento de padrões, cuja eficiência depende da seleção de seus parâmetros: constante de regularização C, função kernel e seus respectivos parâmetros. A escolha equivocada dessas variáveis impacta diretamente na performance do algoritmo, acarretando em fenômenos indesejáveis como o *overfitting* e o *underfitting.* O problema que estuda a procura de parâmetros ótimos para o SVC, em relação às suas medidas de desempenho, é denominado seleção de modelos do SVC. Em virtude do amplo domínio de convergência do kernel gaussiano, a maioria dos métodos destinados a solucionar esse problema concentra-se na seleção da constante C e do parâmetro γ do kernel gaussiano. Dentre esses métodos, a busca por *grid* é um dos de maior destaque devido à sua simplicidade e bons resultados. Contudo, por avaliar todas as combinações de parâmetros (C, γ) dentre o seu espaço de busca, a mesma necessita de muito tempo de processamento, tornando-se impraticável para avaliação de grandes conjuntos de dados. Desta forma, o objetivo deste trabalho é propor um método de seleção de parâmetros do SVC, usando o kernel gaussiano, que combine a técnica *quadtree* à busca por *grid,* para reduzir o número de operações efetuadas pelo *grid* e diminuir o seu custo computacional. A ideia fundamental é empregar a *quadtree* para desenhar a boa região de parâmetros, evitando avaliações desnecessárias de parâmetros situados nas áreas de *underfitting* e *overfitting.* Para isso, desenvolveu-se o método *grid-quadtree* (GQ), utilizando-se a linguagem de programação VB.net em conjunto com os softwares da biblioteca LIBSVM. Na execução do GQ, realizou-se o balanceamento da *quadtree* e criou-se um procedimento denominado refinamento, que permitiu delinear a curva de erro de generalização de parâmetros. Para validar o método proposto, empregaram-se vinte bases de dados referência na área de classificação, as quais foram separadas em dois grupos. Os resultados obtidos pelo GQ foram comparados com os da tradicional busca por *grid* (BG) levando-se em conta o número de operações executadas por ambos os métodos, a taxa de validação cruzada (VC) e o número de vetores suporte (VS) associados aos parâmetros encontrados e a acurácia do SVC na predição dos conjuntos de teste. A partir das análises realizadas, constatou-se que o GQ foi capaz de encontrar parâmetros de excelente qualidade, com altas taxas VC e baixas quantidades de VS, reduzindo em média, pelo menos, 78,8124% das operações da BG para o grupo 1 de dados e de 71,7172% a 88,7052% para o grupo 2. Essa diminuição na quantidade de cálculos efetuados pelo *quadtree* resultou em uma economia de horas de processamento. Além disso, em 11 das 20 bases estudadas a acurácia do SVC-GQ foi superior à do SVC-BG e para quatro delas igual. Isso mostra que o GQ é capaz de encontrar parâmetros melhores ou tão bons quanto os da BG executando muito menos operações.

Palavras-chave: Seleção de modelos do SVC. Kernel gaussiano. *Quadtree.* Redução de operações.

ABSTRACT

The Support Vector Classification (SVC) algorithm is a pattern recognition technique, whose efficiency depends on its parameters selection: the penalty constant C, the kernel function and its own parameters. A wrong choice of these variables values directly impacts on the algorithm performance, leading to undesirable phenomena such as the overfitting and the underfitting. The task of searching for optimal parameters with respect to performance measures is called SVC model selection problem. Due to the Gaussian kernel wide convergence domain, many model selection approaches focus in determine the constant C and the Gaussian kernel γ parameter. Among these, the grid search is one of the highlights due to its easiest way and high performance. However, since it evaluates all parameters combinations (C, γ) on the search space, it requires high computational time and becomes impractical for large data sets evaluation. Thus, the aim of this thesis is to propose a SVC model selection method, using the Gaussian kernel, which integrates the quadtree technique with the grid search to reduce the number of operations performed by the grid and its computational cost. The main idea of this study is to use the quadtree to determine the good parameters region, neglecting the evaluation of unnecessary parameters located in the underfitting and the overfitting areas. In this regard, it was developed the grid-quadtree (GQ) method, which was implemented on VB.net development environment and that also uses the software of the LIBSVM library. In the GQ execution, it was considered the balanced quadtree and it was created a refinement procedure, that allowed to delineate the parameters generalization error curve. In order to validate the proposed method, twenty benchmark classification data set were used, which were separated into two groups. The results obtained via GQ were compared with the traditional grid search (GS) ones, considering the number of operations performed by both methods, the cross-validation rate (CV) and the number of support vectors (SV) associated to the selected parameters, and the SVC accuracy in the test set. Based on this analyzes, it was concluded that GQ was able to find excellent parameters, with high CV rates and few SV, achieving an average reduction of at least 78,8124% on GS operations for group 1 data and from 71,7172% to 88,7052% for group 2. The decrease in the amount of calculations performed by the quadtree lead to savings on the computational time. Furthermore, the SVC-GQ accuracy was superior than SVC-GS in 11 of the 20 studied bases and equal in four of them. These results demonstrate that GQ is able to find better or as good as parameters than BG, but executing much less operations.

Key words: SVC Model Selection. Gaussian kernel. *Quadtree.* Reduction Operations.

LISTA DE FIGURAS

FIGURA 1 - CONJUNTO DE DADOS LINEARMENTE SEPARÁVEL 25
FIGURA 2 - MANEIRAS DE SEPARAR UM CONJUNTO DE DADOS LINEARMENTE SEPARÁVEL 26
FIGURA 3 - SEPARAÇÃO ÓTIMA DE DOIS CONJUNTOS LINEARMENTE SEPARÁVEIS 27
FIGURA 4 - VETORES SUPORTE 30
FIGURA 5 - MAPEAMENTO NÃO LINEAR DO ESPAÇO DE ENTRADA PARA O ESPAÇO DE CARACTERÍSTICAS 31
FIGURA 6 - POSSIBILIDADES DE VIOLAÇÃO DAS MARGENS DE CLASSIFICAÇÃO 36
FIGURA 7 - RELAÇÃO ENTRE A COMPLEXIDADE DE UM MODELO E A SUA TAXA DE ERROS.....41
FIGURA 8 - A GENERALIZAÇÃO DE DIFERENTES HIPERPLANOS SEPARADORES.....42
FIGURA 9 - SEPARAÇÃO ADEQUADA VERSUS *OVERFITTING*.....43
FIGURA 10 - ESPAÇO DE BUSCA DOS PARÂMETROS (C, σ^2) E A EXISTÊNCIA DA BOA REGIÃO.....45
FIGURA 11 - DISTRIBUIÇÃO DAS ÁREAS DO PLANO C – σ^2 46
FIGURA 12 - REPRESENTAÇÃO DA BUSCA POR *GRID*.....48
FIGURA 13 - "*ROUGH LINE*" CORTANDO A BOA REGIÃO.....50
FIGURA 14 - REPRESENTAÇÃO DE UMA ÁRVORE ENRAIZADA 64
FIGURA 15 - EXEMPLO DE UMA ÁRVORE COM ÊNFASE NA PROFUNDIDADE DE SEUS NÓS.....65
FIGURA 16 - REGIÃO E SUA CORRESPONDENTE MATRIZ BINÁRIA 66
FIGURA 17 - FUNCIONAMENTO DA *QUADTREE* 67
FIGURA 18 - SISTEMA DE NUMERAÇÃO ADOTADO 67
FIGURA 19 - ÁRVORE *QUADTREE*.....68
FIGURA 20 - POSSÍVEIS VIZINHOS DE UM QUADRANTE.....69
FIGURA 21 - *QUADTREE* NÃO BALANCEADA VERSUS BALANCEADA 70
FIGURA 22 - ÁRVORE *QUADTREE* NÃO BALANCEADA.....71
FIGURA 23 - ÁRVORE *QUADTREE* BALANCEADA.....72

FIGURA 24 - EXEMPLO DE *QUADTREE* PARA ELUCIDAR O USO DAS FUNÇÕES E OPERAÇÕES 73
FIGURA 25 - FIGURA ORIGINAL E FIGURA COM AS LINHAS "*STROKES*" 79
FIGURA 26 - RESULTADOS DO *IMAGE MATTING* DE HOSAKA, KOBAYASHI E OTSU (2009) 80
FIGURA 27 - LOCALIZAÇÃO DE DEFEITO EM TECIDO 80
FIGURA 28 - EVOLUÇÃO DA *QUADTREE* NA AVALIAÇÃO DE IRRADIÂNCIA SOLAR 81
FIGURA 29 - EXEMPLO DE EXECUÇÃO DA BUSCA POR *GRID* 84
FIGURA 30 - EXEMPLO DE EXECUÇÃO DO MÉTODO *GRID-QUADTREE* 85
FIGURA 31 - EXEMPLO DE NÓ OU QUADRANTE 89
FIGURA 32 - CLASSIFICAÇÕES DE UM QUADRANTE 90
FIGURA 33 - NÓ CINZA ENTENDIDO PELA *QUADTREE* COMO BRANCO 91
FIGURA 34 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DA BG E GQ SEM BALANCEAMENTO PARA O CONJUNTO *IONOSPHERE* 92
FIGURA 35 - SOLUÇÃO GRÁFICA NUMERADA DO *GRID-QUADTREE* SEM BALANCEAMENTO 92
FIGURA 36 - SOLUÇÃO GRÁFICA DO GQ COM BALANCEAMENTO 93
FIGURA 37 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DO GQ SEM E COM BALANCEAMENTO 94
FIGURA 38 - DESTAQUE DOS SINAIS DO QUADRANTE 107 95
FIGURA 39 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DA BG E GQ APÓS REFINAMENTO 95
FIGURA 40 – FLUXOGRAMA DO PROCESSO DE FUNCIONAMENTO DO MÉTODO GQ 97
FIGURA 41 - PONTO DE REFERÊNCIA E DIVISÃO INICIAL DO MÉTODO GQ 99
FIGURA 42 - EXEMPLO DE PONTOS GERADOS ALEATORIAMENTE E PERTENCENTES À MALHA 100
FIGURA 43 - EXEMPLO DE FRAGAMENTAÇÃO IRREGULAR 111
FIGURA 44 - SOLUÇÃO GRÁFICA DO CONJUNTO *MUSHROOM* A16 FORNECIDA PELA BG E GQ 122
FIGURA 45 - SEPARAÇÃO DAS CLASSES DO CONJUNTO A9 DA BASE *CIRCLE AND SQUARE* 134

LISTA DE GRÁFICOS

GRÁFICO 1 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *LIVER DISORDERS* 128
GRÁFICO 2 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *IONOSPHERE* 154
GRÁFICO 3 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *BREAST CANCER* 159
GRÁFICO 4 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *DIABETES* 164
GRÁFICO 5 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *CIRCLE AND SQUARE* 169
GRÁFICO 6 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *IRIS* 174
GRÁFICO 7 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A SVMGUIDE 2 179
GRÁFICO 8 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *VEHICLE* 184
GRÁFICO 9 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *SEGMENT* 189

LISTA DE QUADROS

QUADRO 1 - FUNÇÕES KERNEL MAIS EMPREGADAS NO SVC 33
QUADRO 2 - PSEUDOCÓDIGO DA BUSCA POR *GRID* 48
QUADRO 3 - PSEUDOCÓDIGO DA BUSCA LINEAR 49
QUADRO 4 - PSEUDOCÓDIGO DO MÉTODO DE VAREWYCK e MARTENS (2011) 59
QUADRO 5 - PSEUDOCÓDIGO DO MÉTODO DE WANG, HUANG e CHENG (2014) 60
QUADRO 6 - PSEUDOCÓDIGO DO ALGORITMO DE BALANCEAMENTO DA *QUADTREE* 70
QUADRO 7 - FUNÇÕES NECESSÁRIAS AO ALGORITMO DO VIZINHO 73
QUADRO 8 - OPERAÇÕES UTILIZADAS NO ALGORITMO DO VIZINHO 74
QUADRO 9 - MATRIZ DE ADJACÊNCIA 75
QUADRO 10 - MATRIZ DE REFLEXÃO 75
QUADRO 11 - MATRIZ DE ARESTA COMUM 76
QUADRO 12 - MATRIZ DE OPOSIÇÃO 76
QUADRO 13 - ALGORITMO PARA ENCONTRAR VIZINHOS DE ARESTA 77
QUADRO 14 - ALGORITMO PARA ENCONTRAR VIZINHOS DE VÉRTICE 78
QUADRO 15 - VALIDAÇÃO CRUZADA APLICADA NA SELEÇÃO DE PARÂMETROS DO SVC 87
QUADRO 16 - PSEUDOCÓDIGO DO ALGORITMO DE REFINAMENTO DA *QUADTREE* 96
QUADRO 17 - PSEUDOCÓDIGO DO ALGORITMO DE DETERMINAÇÃO DA SOLUÇÃO INICIAL 100
QUADRO 18 - BASES DE DADOS EMPREGADAS NA DETERMINAÇÃO DE P.. 102
QUADRO 19 - PSEUDOCÓDIGO DO MÉTODO *GRID-QUADTREE* 109
QUADRO 20 - BASES DE DADOS PERTENCENTES AO GRUPO 1 113
QUADRO 21 - TAMANHO DOS CONJUNTOS DE DADOS GERADOS ALEATORIAMENTE 114
QUADRO 22 - BASES DE DADOS PERTENCENTES AO GRUPO 2 115
QUADRO 23 - MATRIZ DE CONFUSÃO GENÉRICA 116
QUADRO 24 - RESUMO DAS ETAPAS DE VALIDAÇÃO DO MÉTODO *GRID-QUADTREE* 116

QUADRO 25 - MATRIZ DE CONFUSÃO DA BG E GQ PARA OS CONJUNTOS *MUSHROOM* A1 A A30 123
QUADRO 26 - MATRIZ DE CONFUSÃO DA BG E GQ PARA O CONJUNTO *LIVER DISORDER*S A10 129
QUADRO 27 - MATRIZ DE CONFUSÃO DA BG E GQ PARA O CONJUNTO *LIVER DISORDERS* A17 129
QUADRO 28 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS EMPREGADAS NO ESTUDO ESTATÍSTICO DE P 190
QUADRO 29 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS DO GRUPO 1 191
QUADRO 30 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS DO GRUPO 2 193

LISTA DE TABELAS

TABELA 1 - IC DO NÚMERO DE OPERAÇÕES (%) REALIZADAS PELO GQ EM FUNÇÃO DE P 103
TABELA 2 - NÚMERO MÉDIO DE OPERAÇÕES (%) EM FUNÇÃO DE P 103
TABELA 3 - DESVIO PADRÃO DO NÚMERO DE OPERAÇÕES (%) EM FUNÇÃO DE P 104
TABELA 4 - IC DA TAXA DE VALIDAÇÃO CRUZADA (%) EM FUNÇÃO DE P 104
TABELA 5 - TAXA MÉDIA DE VC (%) PARA P=60 E P=80 E A RESPECTIVA VARIAÇÃO (%) 105
TABELA 6 - DESVIO PADRÃO DA TAXA DE VALIDAÇÃO CRUZADA (%) EM FUNÇÃO DE P 105
TABELA 7 - IC DA QUANTIDADE DE VETORES SUPORTE (%) EM FUNÇÃO DE P 106
TABELA 8 - QUANTIDADE MÉDIA DE VS (%) PARA P=60 E P=80 E A RESPECTIVA VARIAÇÃO (%) 106
TABELA 9 - DESVIO PADRÃO DA QUANTIDADE DE VETORES SUPORTE (%) EM FUNÇÃO DE P 106
TABELA 10 - VARIAÇÃO (%) DO NÚMERO MÉDIO DE OPERAÇÕES ENTRE P=60 E P=80 107
TABELA 11 - VARIAÇÃO (%) PARA AS MEDIDAS DE DESEMPENHO ENTRE P=60 E P=80 107
TABELA 12 - PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *MUSHROOM* 119
TABELA 13 - COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *MUSHROOM* 121
TABELA 14 - QA E ACURÁCIA DO SVC PARA O CONJUNTO *MUSHROOM* 123
TABELA 15 - PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *LIVER-DISORDERS* 124
TABELA 16 - COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *LIVER-DISORDERS* 126

TABELA 17 - QA E ACURÁCIA DO SVC PARA O CONJUNTO *LIVER-DISORDERS* 127
TABELA 18 – RESUMO ESTATÍSTICO DA TAXA VC (%) DOS PARÂMETROS DETERMINADOS PELA BG E PELO GQ 130
TABELA 19 – RESUMO ESTATÍSTICO DO NÚMERO DE VS (%) DOS PARÂMETROS DETERMINADOS PELA BG E PELO GQ 131
TABELA 20 – RESUMO ESTATÍSTICO DA ACURÁCIA (%) OBTIDA PELOS MÉTODOS SVC-BG E SVC-GQ 132
TABELA 21 – RESUMO ESTATÍSTICO DA REDUÇÃO DE OPERAÇÕES (%) PROPORCIONADA PELO MÉTODO *GRID-QUADREE* 133
TABELA 22 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA OS CONJUNTOS DO GRUPO 2 135
TABELA 23 – COMPARAÇÃO DO NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA OS CONJUNTOS DO GRUPO 2 136
TABELA 24 – QA E ACURÁCIA DO SVC PARA OS CONJUNTOS DA ETAPA 2 . 137
TABELA 25 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *IONOSPHERE* 150
TABELA 26 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *IONOSPHERE* 152
TABELA 27 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *IONOSPHERE* ... 153
TABELA 28 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *BREAST CANCER* 155
TABELA 29 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *BREAST CANCER* 157
TABELA 30 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *BREAST CANCER* 158
TABELA 31 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *DIABETES* 160
TABELA 32 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *DIABETES* 162

TABELA 33 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *DIABETES* 163
TABELA 34 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *CIRCLE AND SQUARE* 165
TABELA 35 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *CIRCLE AND SQUARE* 167
TABELA 36 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *CIRCLE AND SQUARE* 168
TABELA 37 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *IRIS* 170
TABELA 38 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *IRIS* 172
TABELA 39 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *IRIS* 173
TABELA 40 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *SVMGUIDE 2* 175
TABELA 41 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *SVMGUIDE 2* 177
TABELA 42 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *SVMGUIDE 2* 178
TABELA 43 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *VEHICLE* 180
TABELA 44 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *VEHICLE* 182
TABELA 45 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *VEHICLE* 183
TABELA 46 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *SEGMENT* 185
TABELA 47 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *SEGMENT* 187
TABELA 48 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *SEGMENT* 188

LISTA DE SIGLAS

ACO	-	*Ant Colony Optimization*
AFSA	-	*Artificial Fish Swarm Algorithm*
BG	-	Busca por *Grid*
BT	-	Busca Tabu
CPSO	-	*Chained Particle Swarm Optimization*
CSO	-	*Cat Swarm Optimization*
DAGSVM	-	*Directed Acyclic Graph Support Vector Machine*
DPSO	-	*Dynamic Particle Swarm Optimization*
GA	-	*Genetic Algorithm*
GQ	-	*Grid-Quadtree*
IC	-	Intervalo de confiança
KKT	-	Karush – Kuhn –Tucker
LIBSVM	-	*Library for Support Vector Machines*
MA	-	Meta Aprendizado
MCSO	-	*Modified Cat Swarm Optimization*
PAFSA	-	*Parallel Artificial Fish Swarm Algorithm*
PSO	-	*Particle Swarm Optimization*
QA	-	Quantidade de Acertos
RBF	-	Função base radial (ou gaussiana)
SA	-	*Simulated Annealing*
SI	-	Solução Inicial
SOAP	-	*Simple Object Acess Protocol*
SVC	-	*Support Vector Classification*
SVC-BG	-	SVC ajustado com os parâmetros fornecidos pela busca por *grid*
SVC-GQ	-	SVC ajustado com os parâmetros fornecidos pelo *grid-quadtree*
SVM	-	*Support Vector Machine*
SVR	-	*Support Vector Regression*
VC	-	Validação Cruzada
VS	-	Vetores Suporte
VS*Bound*	-	Vetores Suporte do tipo *Bound* (limitados)
XML	-	*eXtensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO 18

1.1 OBJETIVO GERAL 20

1.2 OBJETIVOS ESPECÍFICOS 20

1.3 RELEVÂNCIA E CONTRIBUIÇÕES 20

1.4 ESTRUTURA DO TRABALHO 21

2 SUPPORT VECTOR CLASSIFICATION 23

2.1 HISTÓRICO 23

2.2 CLASSIFICAÇÃO BINÁRIA 24

2.2.1 Algoritmo de classificação de margens rígidas 25

2.2.2 Algoritmo de classificação de margens flexíveis 35

2.3 CLASSIFICAÇÃO EM MÚLTIPLAS CLASSES 38

2.3.1 Método Um contra um 39

2.4 A INFLUÊNCIA DA SELEÇÃO DE PARÂMETROS 40

2.4.1 A preferência pelo kernel gaussiano e a influência do seu parâmetro no comportamento assintótico do SVC 43

2.5 MODELOS DE SELEÇÃO DE PARÂMETROS 46

2.5.1 Busca por *grid* 47

5.2.2 Demais métodos 49

3 *QUADTREE* 63

3.1 DEFINIÇÕES BÁSICAS 63

3.2 CONCEITOS GERAIS 65

3.3 FUNCIONAMENTO DA *QUADTREE* 66

3.4 *QUADTREE* BALANCEADA 69

3.5 DETERMINAÇÃO DE VIZINHOS 72

3.6 APLICAÇÕES DA *QUADTREE* 79

4 PROCEDIMENTOS METODOLÓGICOS 83

4.1 IDEIA NORTEADORA DO MÉTODO *GRID-QUADTREE* 83
4.2 MEDIDAS DE DESEMPENHO DO MÉTODO *GRID – QUADTREE* 86
4.3 CRITÉRIO DE DIVISÃO DOS QUADRANTES 88
4.4 BALANCEAMENTO 90
4.5 REFINAMENTO 94
4.6 O MÉTODO *GRID-QUADTREE* 96
4.6.1 Características gerais 96
4.6.2 Inicialização do método 98
4.6.3 Determinação da solução inicial 99
4.6.4 Estudo estatístico para definir o número de pontos aleatórios P 101
4.6.5 Critérios de parada 108
4.6.6 Pseudocódigo 109
4.7 DIVISÃO UNIFORME E NÃO UNIFORME 110
4.8 VALIDAÇÃO DO MÉTODO *GRID-QUADTREE* 111
4.8.1 Características da busca por *grid* 111
4.8.2 Dados estudados 112
4.8.2.1 Grupo 1– Bases de dados separadas aleatoriamente 112
4.8.2.2 Grupo 2– Bases de dados originalmente separadas 114
4.8.3 Comparação dos resultados 115
4.9 CARACTERÍSTICAS DO COMPUTADOR E VERSÕES DOS *SOFTWARES* 117
5 ANÁLISE E DISCUSSÃO DOS RESULTADOS 118
5.1 RESULTADOS COMPUTACIONAIS DO GRUPO 1 118
5.2 RESULTADOS COMPUTACIONAIS DO GRUPO 2 135
6 CONSIDERAÇÕES FINAIS 139
6.1 SUGESTÕES PARA TRABALHOS FUTUROS 141
REFERÊNCIAS 143
APÊNDICE 150

1 INTRODUÇÃO

O algoritmo *Support Vector Machine* (SVM), desenvolvido por Vapnik e seus colaboradores (BOSER; GUYON; VAPNIK, 1992), tem apresentado excelentes resultados na área da classificação e da regressão de dados. Na sua versão destinada à classificação, denominada *Support Vector Classification* (SVC), o algoritmo tem por objetivo encontrar um hiperplano separador de máxima margem que classifique os dados de teste corretamente.

Entretanto, o desempenho do SVC depende significativamente da seleção de seus parâmetros: constante de regularização C, função kernel e seus respectivos parâmetros. A escolha inadequada dessas variáveis, além de elevar o tempo de treinamento do SVC, acarreta em fenômenos indesejáveis como o *overfitting* e o *underfitting*, reduzindo a acurácia do algoritmo.

Dentre as possibilidades de função kernel empregadas no SVC, o gaussiano é o que tem maior domínio de convergência e eficiência computacional (PANG *et al.,* 2011; WANG *et al.,*2014), portanto, ele é normalmente a primeira opção dos usuários e, consequentemente, o mais utilizado no algoritmo (HSU, CHANG, LIN, 2010; LIN *et al.* (2008a). Por essa razão, a tarefa de selecionar parâmetros para o SVC frequentemente se limita à procura do par (C, γ), onde γ é o parâmetro do kernel gaussiano.

O problema que estuda a determinação de parâmetros ótimos para o SVC, em relação a suas medidas de desempenho (taxa de validação cruzada, número de vetores suporte, dentre outros), é denominado seleção de modelos do SVC (CHAPELLE; VAPNIK, 1999). Nos últimos anos, muitos foram os métodos desenvolvidos para resolver esse problema, dentre os quais destaca-se a busca por *grid* (BG) por ser um dos mais utilizados.

A preferência pela BG se dá tanto por sua simplicidade de uso (seja por meio de pacotes computacionais disponibilizados na internet e/ou fácil reprodução) quanto pelo fornecimento de bons resultados. Contudo, em virtude da BG analisar todas as combinações de parâmetros (C, γ) em seu espaço de busca, a mesma apresenta alto custo computacional. Logo, a sua aplicação na avaliação de grandes conjuntos de dados torna-se muito demorada e, em muitos dos casos, impraticável. Desta forma, pesquisadores dedicam-se a elaborar métodos de seleção de parâmetros que sejam competitivos à acurácia da BG, porém com uma execução mais rápida.

Keerthi e Lin (2003), ao estudarem o comportamento assintótico do SVC usando o kernel gaussiano, permitiram o entendimento do espaço de busca dos parâmetros C e γ. Os seus resultados mostraram a existência de uma curva de erro de generalização, que separa a área de parâmetros em duas regiões: uma caracterizada pelo *underfitting* e *overfitting* e outra denominada boa região, na qual se encontra o par (C, γ) ótimo.

A partir da pesquisa de Keerthi e Lin (2003) evidencia-se que muitas das combinações de (C, γ), por impactarem negativamente na performance do SVC, são desnecessárias ao processo de avaliação de parâmetros. Dessa maneira, a procura da solução pode-se restringir à boa região, tornando a sua localização e identificação muito importante.

Neste contexto, aponta-se a técnica *quadtree* como uma ferramenta para representar dados (regiões), cujo funcionamento baseia-se na divisão sucessiva do espaço em quadrantes do mesmo tamanho (SAMET, 1981). O conceito fundamental da *quadtree* está em identificar quais dos seus quadrantes estão inteiramente contidos na região de interesse, parcialmente inseridos ou vazios. Somente os quadrantes que estiverem parcialmente contidos, ou seja, que possuírem dados internos e externos à área considerada, são recursivamente divididos até que se tornem homogêneos. Encerra-se o processo de divisão da *quadtree* quando a condição de homogeneidade dos quadrantes é obtida.

Assim sendo, na *quadtree* a região analisada é representada por uma quantidade otimizada de quadrantes, uma vez que as áreas semelhantes são interpretadas como um mesmo nó da árvore. Diante dessa propriedade e das características do espaço de busca dos parâmetros (C, γ), observadas por Keerthi e Lin (2003), surgiu a motivação da presente tese, que é incorporar a técnica *quadtree* à busca por *grid*, de forma a dispensar a varredura completa da malha e reduzir o seu tempo computacional. Com isso, visa-se agregar as qualidades da técnica tradicional, simplicidade e bons resultados, com a capacidade de otimização da *quadtree.*

1.1 OBJETIVO GERAL

O objetivo geral deste trabalho é propor um método de seleção de parâmetros do SVC, usando o kernel gaussiano, que combine a técnica *quadtree* à busca por *grid*, para reduzir o número de operações (cálculos) efetuadas pelo *grid* e diminuir seu custo computacional.

1.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos compreendem:

- Estudar o problema da seleção de modelos do SVC e o funcionamento da técnica *quadtree.*
- Criar um procedimento de refinamento da *quadtree* aplicado à seleção de parâmetros do SVC.
- Desenvolver o método *grid-quadtree* para a determinação de parâmetros (C, γ) do SVC considerando as medidas de desempenho taxa de validação cruzada e número de vetores suporte.
- Implementar computacionalmente os métodos *grid-quadtree* e busca por *grid* usando a linguagem VB.net e os *softwares* da biblioteca LIBSVM.
- Testar o método *grid-quadtree* utilizando bases de dados referência na área da classificação de dados.
- Comparar os resultados do método *grid-quadtree* com os da busca por *grid.*
- Avaliar o desempenho computacional do método *grid-quadtree* para a seleção de parâmetros do SVC.

1.3 RELEVÂNCIA E CONTRIBUIÇÕES

Determinar adequadamente os parâmetros do SVC é uma tarefa essencial para se obter o melhor desempenho do algoritmo. Os modelos de seleção de parâmetros, neste contexto, auxiliam o trabalho de usuários do SVC poupando-os de tomar decisões arbitrárias e que dependam da sua experiência. Contudo, os métodos

disponíveis devem, além de fornecer repostas que propiciem alta acurácia ao SVC, ser rápidos e de fácil empregabilidade.

Desta forma, o método *grid-quadtree* contribui no campo de pesquisa por beneficiar-se das qualidades da busca por *grid*, simplicidade de uso e bons resultados, e apresentar menor custo computacional. A utilização da *quadtree* combinada ao *grid* permite a redução de operações desnecessárias executadas pela técnica tradicional, negligenciando avaliações de parâmetros localizados nas áreas de *underfitting* e *overfitting* do espaço de busca.

Essa redução de operações, além de propiciar a diminuição do tempo computacional em problemas conhecidos, de bases de dados referência, também viabiliza a possível utilização do método proposto em problemas de maiores dimensões. Ainda, por a *quadtree* não ser uma técnica de complexa implementação computacional e dispensar a determinação de parâmetros adicionais, a sua aplicação torna-se acessível aos usuários do SVC.

Por fim, como não foram evidenciados, na revisão de literatura realizada, trabalhos que empreguem a técnica *quadtree* no problema de seleção de parâmetros do SVC, o método *grid-quadtree* contribui para o desenvolvimento da pesquisa da área e fornece uma nova ferramenta de determinação de parâmetros aos utilizadores do SVC.

1.4 ESTRUTURA DO TRABALHO

A presente tese está organizada em seis capítulos. No primeiro, introduz-se o tema do trabalho assim como seus objetivos e contribuições.

No segundo capítulo, apresenta-se a formulação matemática do SVC e explica-se a importância da correta seleção de parâmetros para o desempenho do algoritmo. Na sequência, faz-se uma revisão dos principais métodos de determinação de parâmetros evidenciados na literatura, dando ênfase aos mais recentes.

No terceiro capítulo, expõem-se os conceitos da técnica *quadtree* e os algoritmos necessários para o seu funcionamento. Além disso, mostram-se algumas das suas possíveis aplicações.

No capítulo quatro, explicam-se os procedimentos metodológicos adotados para o desenvolvimento do método *grid-quadtree*, seu pseudocódigo, bases de dados utilizadas e critérios de validação.

No capítulo cinco, faz-se a análise dos resultados obtidos pelo método *grid-quadtree* e os compara com os da busca por *grid.*

Por fim, no capítulo seis realizam-se as considerações finais do trabalho e as sugestões para pesquisas futuras.

2 SUPPORT VECTOR CLASSIFICATION

Neste capítulo, expõem-se a teoria e a formulação matemática do algoritmo *Support Vector Classification* (SVC) e destaca-se a influência da seleção de parâmetros no seu desempenho. Além da explicação de seus conceitos e características fundamentais, faz-se uma revisão de literatura dos principais métodos de determinação de parâmetros aplicados ao SVC, ressaltando desde as pesquisas precursoras até as mais atuais.

2.1 HISTÓRICO

O *Support Vector Machine* (SVM) é uma generalização não linear do algoritmo *Generalized Portrait*, criado na década de sessenta, na Rússia, por Vapnik e Lerner (1963) e Vapnik e Chervonenkis (1964). Entretanto, o surgimento da sua presente versão ocorreu nos anos noventa no *AT&T Bell Laboratories* devido ao trabalho de Vapnik e seus colaboradores.

Desenvolvido por Boser, Guyon e Vapnik, o algoritmo SVM foi apresentado pela primeira vez em 1992 na *Annual Workshop on Computational Learning Theory.* Em sua concepção original, denominada “Algoritmo de treinamento para classificadores de margem ótima”, o modelo visava encontrar um hiperplano separador de máxima margem, que classificasse dados linearmente separáveis tanto no espaço de entrada quanto no espaço de características, sendo a classificação nesse último viabilizada por funções de transformação kernel (BOSER; GUYON; VAPNIK, 1992).

No ano seguinte, publicou-se o trabalho Guyon, Boser e Vapnik (1993), que aprofundava conceitos da pesquisa inicial. No entanto, foi em 1995 que Cortes e Vapnik sugeriram uma modificação importante ao algoritmo então proposto, que foi a introdução de variáveis de folga. Essa alteração possibilitou a determinação de uma superfície de decisão de margens flexíveis, que estendeu a aplicação do algoritmo aos conjuntos de dados não linearmente separáveis (CORTES; VAPNIK, 1995).

A partir do trabalho de Cortes e Vapnik (1995), o método ficou conhecido por *Support Vector Machine*. Todavia, como atualmente o termo SVM abrange tanto as técnicas de regressão, *Support Vector Regression* (SVR), quanto a de classificação, *Support Vector Classification* (SVC), a presente pesquisa utiliza a terminologia SVC

para referir-se ao algoritmo de classificação, que é a formulação de interesse nesta tese.

Em virtude do seu bom desempenho e acurácia nos resultados, verifica-se aplicações do SVC em diversas áreas do conhecimento, tais como: na biometria e segurança, medicina, indústria e no mercado financeiro. São alguns exemplos da sua utilização: a verificação e o reconhecimento de assinaturas (ÖZGÜNDÜZ; SENTÜRK; KARSLIGIL, 2005), a classificação de digitais (YAO *et al.*, 2001), a detecção de faces (OSUNA; FREUND; GIROSI, 1997); os diagnósticos de câncer e de outras doenças (MOCELLIN *et al.*, 2006; AKAY, 2009; JIANG; ZOU, 2013; LEBRUN *et al,* 2008, JIANG; MISSOUM; CHEN, 2014); a identificação de falhas em máquinas rotativas (ZHANG *et al.,*2015), a avaliação de compostos e propriedades físicas de materiais (DEVOS *et al.,* 2009) e a análise de crédito bancário (ALES *et al.,* 2009).

2.2 CLASSIFICAÇÃO BINÁRIA

O algoritmo SVC foi originalmente desenvolvido para a classificação binária, mas o mesmo pode ser empregado para a classificação em múltiplas classes. As explicações referentes ao SVC serão iniciadas para a primeira situação e posteriormente estendidas para a segunda.

O modelo mais simples do SVC e o primeiro a ser introduzido por Boser, Guyon e Vapnik (1992) foi o algoritmo de margens rígidas. Porém, por trabalhar apenas com dados linearmente separáveis, ele não atende a muitas das aplicações reais. Contudo, devido aos seus conceitos constituírem a base do SVC, sua descrição é crucial para o entendimento do modelo mais sofisticado e completo: o de margens flexíveis (CRISTIANINI; SHAWE – TAYLOR, 2006), que é a sua atual versão.

Desta forma, inicia-se a explicação da teoria do SVC pelo algoritmo de margens rígidas e, na sequência, apresenta-se o de margens flexíveis, que permite a classificação de dados não linearmente separáveis e a sua extensão a casos práticos. Assim, a explanação do SVC é realizada na mesma ordem em que ele foi desenvolvido por Vapnik e seus colaboradores. Ressalta-se ainda que as notações utilizadas neste capítulo se baseiam em Cortes e Vapnik (1995), Vapnik (1999) e Cristianini e Shawe- Taylor (2006), livro referência na área do SVM.

2.2.1 Algoritmo de classificação de margens rígidas

O funcionamento do SVC baseia-se na metodologia do aprendizado supervisionado, que faz com que o sistema aprenda a partir de uma amostra de dados conhecida. Tal processo ocorre em duas etapas: a de treinamento e a de teste.

No treinamento, fornecem-se pares de entrada e saída $\{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$ ao sistema, com o intuito de ajustar os parâmetros livres e determinar uma função que relacione os respectivos valores de $\boldsymbol{x}_i$ e y_i. Para isso, penalizam-se os erros de estimação e valorizam-se os acertos de modo que o algoritmo retenha informações relevantes e adquira conhecimento. Já na fase de teste, emprega-se a função estabelecida no treinamento para gerar as saídas dos padrões desconhecidos.

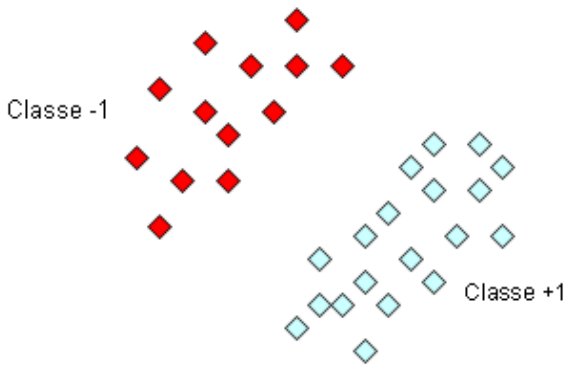
De acordo com Boser, Guyon e Vapnik (1992), os dados de treinamento são representados por l exemplos $\boldsymbol{x}_i$ com rótulos y_i:

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3), \dots, (\boldsymbol{x}_l, y_l) \tag{1}$$

$$onde \begin{cases} y_i = 1 & se\ \boldsymbol{x}_i \in classe\ A \\ y_i = -1 & se\ \boldsymbol{x}_i \in classe\ B. \end{cases}$$

A partir do conjunto de treinamento (1), composto por duas classes linearmente separáveis, o objetivo do SVC é encontrar uma função de boa capacidade de generalização capaz de classificar os dados de teste corretamente. A FIGURA 1 ilustra um exemplo para o conjunto (1).

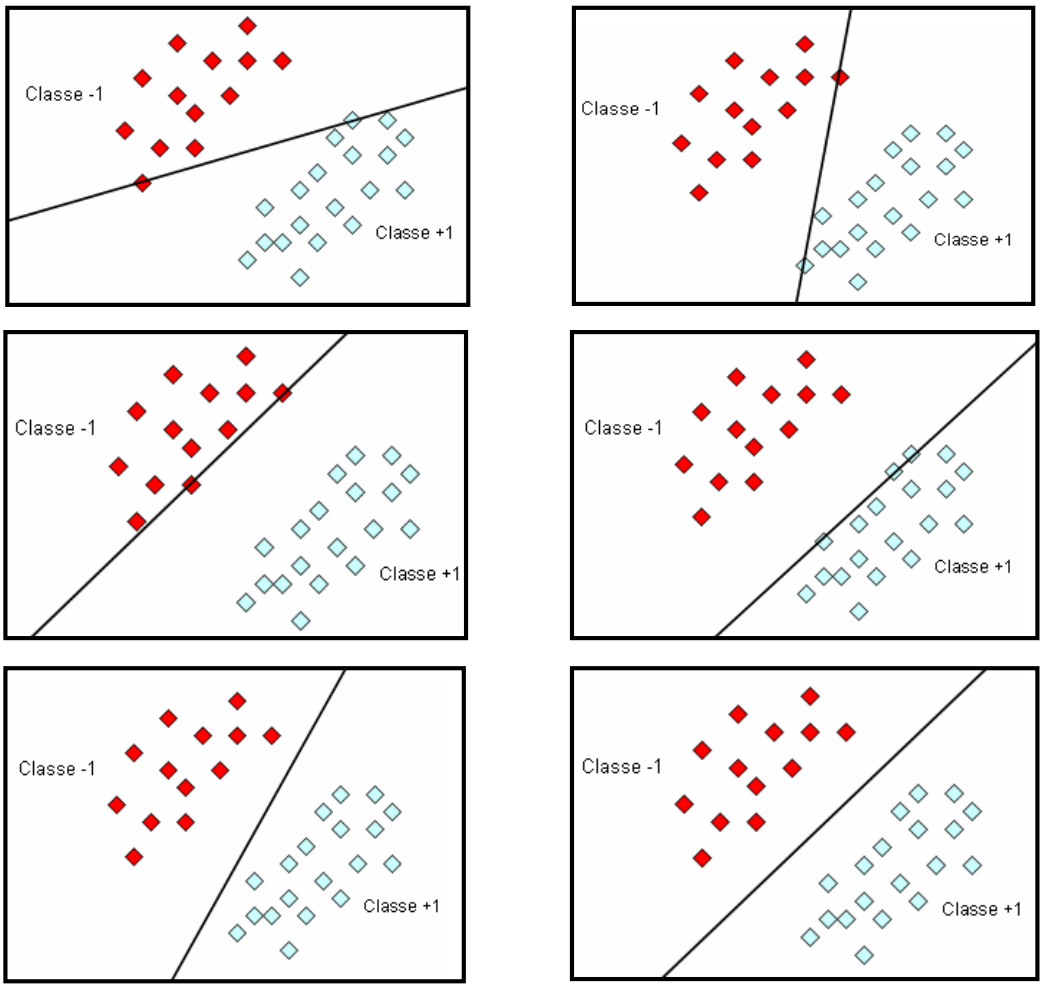
FIGURA 1 - CONJUNTO DE DADOS LINEARMENTE SEPARÁVEL



FONTE: ALES (2008).

A FIGURA 2 mostra que existem diversos meios de separar sem erros o conjunto da FIGURA 1. Entretanto, há somente uma maneira na qual se maximiza a margem de classificação, que é caracterizada pela distância entre o hiperplano separador e o vetor mais próximo de cada classe. Essa situação peculiar é observada no canto inferior direito da FIGURA 2.

FIGURA 2 - MANEIRAS DE SEPARAR UM CONJUNTO DE DADOS LINEARMENTE SEPARÁVEL



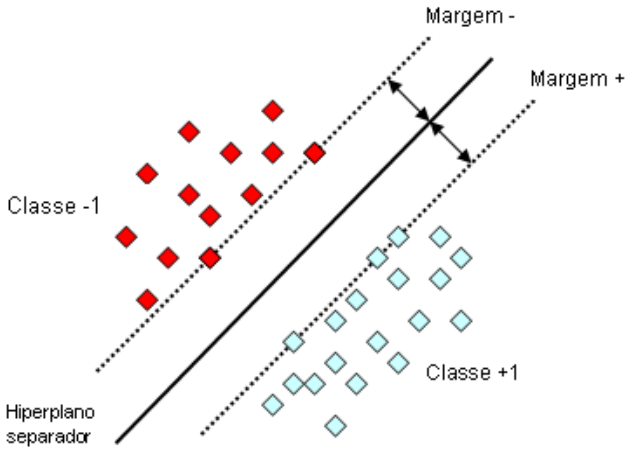
FONTE: ALES (2008).

O hiperplano que maximiza a distância entre as duas margens de classificação é denominado hiperplano ótimo. Resultados teóricos descritos por Vapnik (1995) mostram que a maximização das margens implica em uma maior generalização do algoritmo de aprendizagem. Desta forma, a finalidade do SVC é determinar esse hiperplano particular, ilustrado em detalhes na FIGURA 3, definido por

$$f(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b \tag{2}$$

onde: $f: \mathbb{R}^n \rightarrow \mathbb{R}$; $\boldsymbol{w}$ é o vetor dos pesos e b é o *bias.*

FIGURA 3 - SEPARAÇÃO ÓTIMA DE DOIS CONJUNTOS LINEARMENTE SEPARÁVEIS



FONTE: ALES (2008).

Segundo Boser, Guyon e Vapnik (1992), maximizar a margem de classificação equivale a minimizar a norma $\|\boldsymbol{w}\|$. Considerando que as margens positiva e negativa são representadas respectivamente por (3) e (4)

$$\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b = 1 \tag{3}$$

$$\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b = -1, \tag{4}$$

um vetor de treinamento está corretamente classificado quando satisfaz as expressões

$$\begin{aligned} \langle \boldsymbol{w} \cdot \ \boldsymbol{x}_i \rangle + b \leq -1, \ \ se \ \ y_i = -1 \\ \langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b \geq +1, \ \ se \ \ y_i = +1. \end{aligned} \tag{5}$$

Reescrevendo as duas inequações de (5) em uma única relação, tem-se que:

$$y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1, \quad i = 1, \ldots, l. \tag{6}$$

Logo, o problema de otimização primal do SVC que determina o hiperplano separador de margens máximas, formulado por Boser, Guyon e Vapnik (1992), é dado por

$$\begin{aligned} &minimizar \quad \|\boldsymbol{w}\|^2 \\ &sujeito\ a\colon \ y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1, \qquad i = 1, \dots, l \end{aligned} \tag{7}$$

onde: $\boldsymbol{w} \in \mathbb{R}^n$ e $b \in \mathbb{R}$ são as incógnitas.

De acordo com Vapnik (1999), solucionar o problema de otimização (7) é equivalente a encontrar o ponto de sela da função Lagrangeana[1]

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + \ b) - 1] \tag{8}$$

onde: $\alpha_i \geq 0$ são os multiplicadores de Lagrange. Para isso, a função (8) deve ser minimizada em relação às variáveis primais $\boldsymbol{w}$, b e maximizada em relação a $\alpha_i \geq 0$.

No ponto de sela, as soluções ótimas $(\boldsymbol{w}^*, b^*)$, $\boldsymbol{\alpha}^*$ devem satisfazer as condições (9) e (10)

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = 0 \tag{9}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} = \ 0 \tag{10}$$

da onde se obtém propriedades importantes relacionadas ao hiperplano ótimo

$$\boldsymbol{w}^* = \sum_{i=1}^{l} y_i \alpha_i^* \boldsymbol{x}_i \tag{11}$$

$$\sum_{i=1}^{l} y_i \alpha_i^* = 0\,. \tag{12}$$

[1] Boser, Guyon e Vapnik (1992) explicam que o fator $\frac{1}{2}$ foi inserido na equação (8) por questões estéticas e que o mesmo não afeta a solução.

Substituindo as expressões (11) e (12) na função (8), chega-se à:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle \,. \tag{13}$$

A obtenção da função (13) e o uso dos conceitos da teoria de otimização possibilitam introduzir a representação dual do SVC, que é dada por:

$$\begin{aligned} & maximizar \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i \cdot \boldsymbol{x}_j \rangle \\ & sujeito \;\; a\text{:} \;\; \sum_{i=1}^{l} y_i \alpha_i = 0 \\ & \qquad \alpha_i \geq 0, \qquad i = 1, \ldots, l. \end{aligned} \tag{14}$$

O teorema de Kuhn - Tucker e as suas relações, em especial as condições de complementaridade de Karush – Kuhn –Tucker (KKT), representadas por (18), se aplicam ao SVC e fornecem informações muito úteis quanto à estrutura da sua solução (CRISTIANINI; SHAWE – TAYLOR, 2006).

Teorema de Kuhn e Tucker: *Dado um problema de otimização com domínio convexo* $\boldsymbol{\Omega} \subseteq \mathbb{R}^n$,

$$\begin{aligned} & minimizar \quad f(\boldsymbol{w}, b), \qquad \boldsymbol{w} \in \boldsymbol{\Omega}, \\ & \text{s}ujeito\ a\text{:} \quad g_i(\boldsymbol{w}, b) \leq 0, \qquad i = 1, \ldots, l, \end{aligned} \tag{15}$$

com $f \in C^1$ *convexa e* g_i *funções afins, a condição necessária e suficiente para que o par* $(\boldsymbol{w}^*, b^*)$ *seja ótimo é a existência de* $\boldsymbol{\alpha}^*$, *satisfazendo*

$$\frac{\partial L(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}^*)}{\partial \boldsymbol{w}} = \mathbf{0}, \tag{16}$$

$$\frac{\partial L(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}^*)}{\partial b} = 0, \tag{17}$$

$$\alpha_i{}^* g_i(\boldsymbol{w}^*, b^*) = 0, \quad i = 1, \dots, l \tag{18}$$

$$g_i(\boldsymbol{w}^*, b^*) \leq 0, \quad i = 1, \dots, l \tag{19}$$

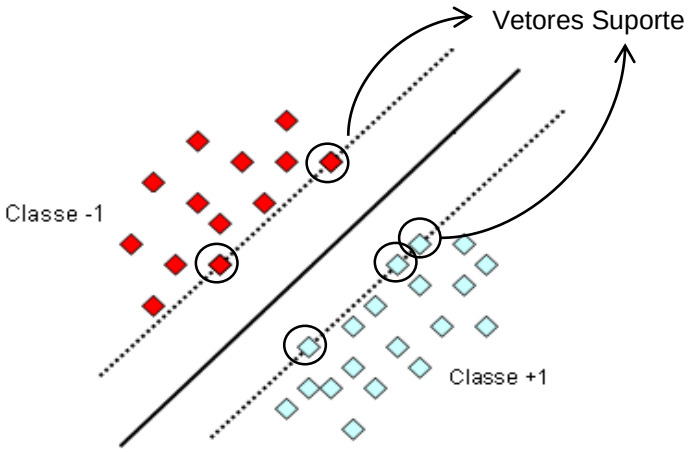
$$\alpha_i{}^* \geq 0, \quad i = 1, \dots, l. \tag{20}$$

Das condições de complementaridade de KKT, equação (18), tem-se que as soluções ótimas $\boldsymbol{\alpha}^*, (\boldsymbol{w}^*, b^*)$ devem satisfazer

$$\alpha_i^* [y_i(\langle \boldsymbol{w}^* \cdot \boldsymbol{x}_i \rangle + b^*) - 1] = 0, \qquad i = 1, \dots, l. \tag{21}$$

A expansão (21) implica que apenas os dados de entrada situados sobre as margens de classificação, ou seja, os que atendem à $y_i(\langle \boldsymbol{w}^* \cdot \boldsymbol{x}_i \rangle + b^*) = 1$, têm seu respectivo multiplicador de Lagrange $\alpha_i^* \neq 0$, os demais possuem $\alpha_i^* = 0$. Pela equação (11), percebe-se que somente os pontos com $\alpha_i^* \neq 0$ são necessários para calcular o vetor dos pesos $\boldsymbol{w}^*$. Portanto, por serem os únicos a influenciar a construção do hiperplano separador, eles são chamados de vetores suporte (VS). A FIGURA 4 destaca os referidos vetores.

FIGURA 4 - VETORES SUPORTE



FONTE: BELTRAMI (2009).

Assim, tem-se que os multiplicadores de Lagrange α_i^* quantificam o quão importante cada padrão de treinamento será para a solução final, de forma que, na

representação dual, o hiperplano ótimo pode ser expresso unicamente em termos de VS

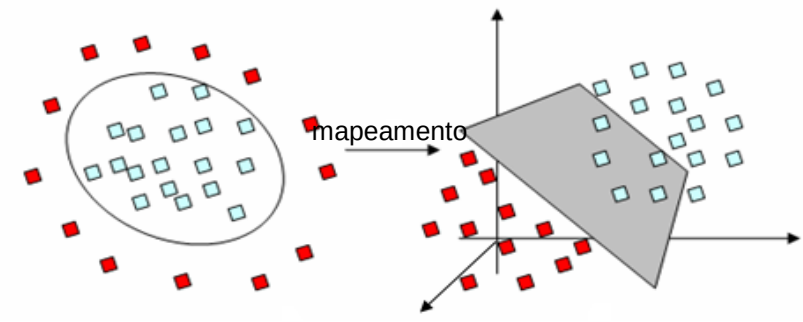
$$f(\boldsymbol{x}, \boldsymbol{\alpha}^*, b^*) = \sum_{i \in VS} y_i \alpha_i^* \langle \boldsymbol{x}_i \cdot \boldsymbol{x} \rangle + b^*. \tag{22}$$

Segundo Boser, Guyon e Vapnik (1992), a dualidade do SVC é explorada para melhorar a eficácia e flexibilidade do algoritmo. A resolução do SVC via problema primal, além de ser difícil devido às suas restrições de desigualdade, não fornece informações acerca dos vetores suporte. Ademais, visto que o número de VS é normalmente menor que a quantidade total de dados de treinamento, trabalhar com uma função restrita aos VS, como a (22), resulta em maior eficiência computacional (GUYON; BOSER; VAPNIK, 1993).

Até o momento, considerou-se que a separação dos dados (tanto via primal quanto dual) foi sempre realizada no espaço de entrada por meio de funções lineares. No entanto, essa suposição é raramente conveniente, uma vez ela não atende às peculiaridades dos casos reais.

Nesse sentido, uma das grandes vantagens da representação dual é que ela permite o mapeamento não linear dos dados de entrada para um espaço de maior dimensão, denominado espaço de características, onde a separação linear torna-se possível. A FIGURA 5 ilustra essa situação, da qual repara-se que os dados não linearmente separáveis no espaço de entrada (bidimensional) assim os tornam quando mapeados para o tridimensional.

FIGURA 5 - MAPEAMENTO NÃO LINEAR DO ESPAÇO DE ENTRADA PARA O ESPAÇO DE CARACTERÍSTICAS



FONTE: CARVALHO (2005).

Normalmente, tal procedimento é executado alterando-se a representação dos dados conforme segue:

$$\boldsymbol{x} = (x_1, \dots, x_n) \mapsto \boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \dots, \phi_N(\boldsymbol{x})), \tag{23}$$

onde $\boldsymbol{\phi}: X \rightarrow F$, é um mapeamento não linear do espaço de entrada X para o espaço de características F.

Para implementar (23) no SVC, basta substituir na função objetivo de (14) as entradas $\boldsymbol{x}$ por $\boldsymbol{\phi}(\boldsymbol{x})$**,** onde $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^N$, com $N \gg n$. Mas, devido ao custo computacional de se realizar o produto interno $\langle \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle$ diretamente no espaço de características, opta-se por efetuar um mapeamento implícito por meio das funções kernel, que dependem unicamente das variáveis do espaço de entrada (CHAVES, 2006).

Definição de kernel: *Um kernel é uma função K, tal que para todo $\boldsymbol{x}, \boldsymbol{z} \in X$*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \langle \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{z}) \rangle, \tag{24}$$

onde ϕ é um mapeamento de X para um espaço característico F.

O teorema de Mercer, apresentado a seguir, estabelece as condições para que uma função $K(\boldsymbol{x}, \boldsymbol{z})$ seja um kernel.

Teorema de Mercer: *Seja X um conjunto compacto de $\mathbb{R}^n$ e K uma função contínua e simétrica. Existe um mapeamento ϕ e uma expansão da forma*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{z}) \tag{25}$$

associados a autovalores $\lambda_i \geq 0$, se e somente se, para todo $f(\boldsymbol{x})$ no qual

$$\int_X f(\boldsymbol{x})^2 d\boldsymbol{x} < \infty \tag{26}$$

implica que:

$$\int_X \int_X K(\boldsymbol{x},\boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z} \geq 0 \tag{27}$$

A partir das condições de Mercer tem-se a seguinte proposição:

Proposição: *Seja X um espaço de entrada finito com* $K(\boldsymbol{x},\boldsymbol{z})$ *uma função simétrica em X. Então* $K(\boldsymbol{x},\boldsymbol{z})$ *é uma função kernel se e somente se, a matriz*

$$\mathbf{K} = \left(K(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)_{i,j=1}^{n} \tag{28}$$

é semi definida positiva, ou seja, seus autovalores são não negativos.

O QUADRO 1 destaca os exemplos mais comuns de funções kernel empregadas no SVC. Contudo, em Guyon, Boser e Vapnik (1993) encontram-se outras possibilidades, além das citadas nesta tese.

QUADRO 1 - FUNÇÕES KERNEL MAIS EMPREGADAS NO SVC

TIPO DE KERNEL	**EQUAÇÃO**	**PARÂMETROS**	**Nº EQ.**
Linear	$K(\boldsymbol{x_i},\boldsymbol{x_j}) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$	Não há	(29)
Polinomial homogêneo	$K(\boldsymbol{x}_\mathrm{i},\boldsymbol{x}_\mathrm{j}) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)^\mathrm{p}$	p = grau do polinômio	(30)
Polinomial não homogêneo	$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + k)^p$	p = grau do polinômio; k = constante	(31)
Sigmoidal	$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = tanh(\langle \kappa \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + k)$	κ = coeficiente; k = constante negativa	(32)
Gaussiano (ou função base radial - RBF)	$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}}$	σ	(33)
	$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = e^{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}$	γ	(34)

FONTE: A autora (2016).

Pelo QUADRO 1, observa-se que o kernel gaussiano (ou função base radial) pode ser representado tanto pela equação (33) quanto pela (34). A revisão de literatura deste trabalho faz menção as duas formas, visto que alguns artigos abordam a primeira e outros a segunda. No entanto, nota-se que ambas são essencialmente iguais, já que

$$\gamma = \frac{1}{2\sigma^2}. \tag{35}$$

Introduzindo o conceito de mapeamento implícito no SVC, a sua formulação dual passa a ser

$$\begin{aligned} &maximizar \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \\ &sujeito\ a: \sum_{i=1}^{l} y_i \alpha_i = 0 \\ &\alpha_i \geq 0, \qquad i = 1, \dots, l. \end{aligned} \tag{36}$$

As condições de KKT novamente se aplicam a (36) e fornecem as mesmas informações descritas para (14). Entretanto, o vetor dos pesos, que era facilmente determinado por (11), já não pode mais ser descrito explicitamente, pois seu cálculo, equação (37), envolve a função $\phi(x)$ que nem sempre é conhecida.

$$\boldsymbol{w}^* = \sum_{i=1}^{l} y_i \alpha_i^* \phi(\boldsymbol{x}_i) \tag{37}$$

Por fim, a função de decisão do SVC (hiperplano separador) é dada por

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} y_i \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*, \tag{38}$$

onde $K(\boldsymbol{x}_i, \boldsymbol{x})$ é um kernel escolhido *a priori* e o *bias* b^* é determinado por

$$b^* = -\frac{1}{2}\sum_{i=1}^{l} \alpha_i^* y_i [K(\boldsymbol{x}_i, \boldsymbol{x}_A) + K(\boldsymbol{x}_i, \boldsymbol{x}_B)], \quad (39)$$

em que $\boldsymbol{x}_A$ e $\boldsymbol{x}_B$ são dois vetores suporte arbitrários pertencentes respectivamente às classes A e B, conforme descrito em (1).

Finaliza-se aqui a explicação do “Algoritmo de treinamento para classificadores de margem ótima”, também conhecido por algoritmo de margens rígidas, apresentado por Boser, Guyon e Vapnik (1992). Apesar dessa versão não ser aplicável à maioria dos problemas reais, devido a ela trabalhar somente com dados linearmente separáveis, ela é o ponto de partida para a construção da forma mais completa e atual do SVC.

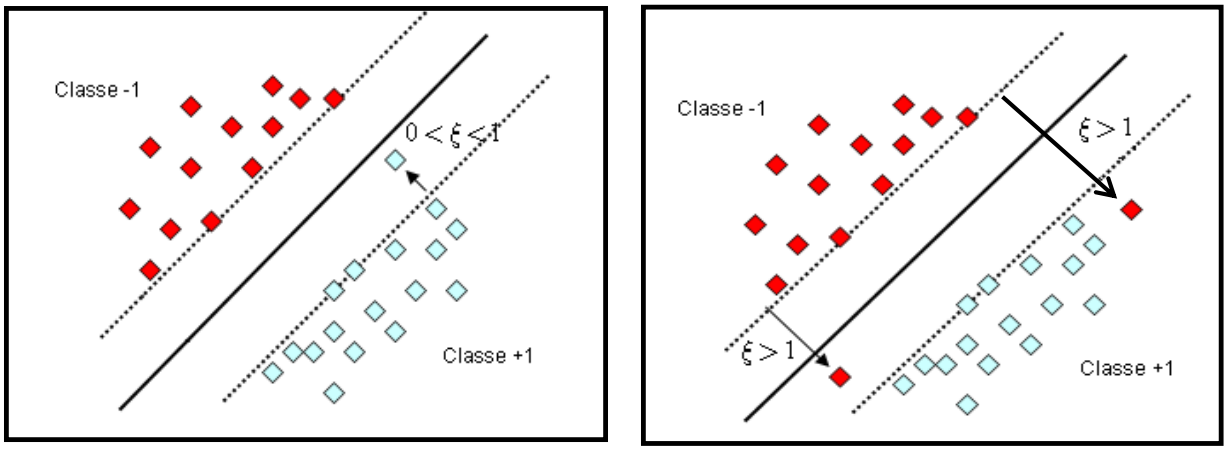
2.2.2 Algoritmo de classificação de margens flexíveis

A principal limitação do algoritmo de margens rígidas, exposto na subseção 2.2.1, é que ele não considera em nenhum momento a existência de erros de classificação. Por essa razão, ele não se aplica a muitos dos exemplos reais (dados não linearmente separáveis), uma vez que nesses conjuntos existem pontos de treinamento que infringem a restrição (6) do SVC (CRISTIANINI; SHAWE – TAYLOR, 2006).

A fim de superar esse entrave, Cortes e Vapnik (1995) introduziram variáveis de folga ξ_i ao SVC, para permitir que as restrições relacionadas às margens pudessem ser violadas. Desta forma, eles associaram variáveis $\xi_i \geq 0$ aos vetores de treinamento para medir os desvios encontrados em relação às condições ideais de separação. Assim, se um vetor i está corretamente separado $\xi_i = 0$.

Pela restrição (6), constata-se que a violação das margens pode ocorrer de duas maneiras distintas, as quais estão ilustradas na FIGURA 6.

FIGURA 6 - POSSIBILIDADES DE VIOLAÇÃO DAS MARGENS DE CLASSIFICAÇÃO



FONTE: BELTRAMI (2009).

Observando-se a FIGURA 6 (à esquerda), repara-se que a primeira possibilidade se dá quando o ponto $\boldsymbol{x}_i$ se localiza dentro da região de separação e no lado correto da classificação. Para essa situação, tem-se que $0 < \xi_i \leq 1$. Já o segundo modo (FIGURA 6 à direita) ocorre quando $\boldsymbol{x}_i$ se situa no lado incorreto da superfície de decisão, podendo estar dentro ou fora da região de separação. Nesse caso, a variável de folga tem valor $\xi_i > 1$.

Feitas essas considerações, diz-se que um vetor de treinamento está corretamente classificado quando ele satisfaz

$$y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i, \qquad i = 1, \dots, l, \tag{40}$$

da onde nota-se que se $\xi_i = 0$, a relação (40) se reduz a (6).

Com isso, o problema de otimização primal do SVC, que determina o hiperplano ótimo de margens flexíveis, torna-se

$$\begin{aligned} &minimizar \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \mathrm{C}\sum_{i=1}^{l} \xi_i \\ &sujeito\ a: \quad y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i \\ &\qquad\qquad \xi_i \geq 0, \qquad i = 1, \dots, l \end{aligned} \tag{41}$$

onde C é denominada constante de regularização, pois pondera os termos da função de minimização.

Analisando a função objetivo de (41), tem-se que o termo $\frac{1}{2}\|w\|^2$ visa maximizar a margem do hiperplano enquanto que $\mathrm{C}\sum_{i=1}^{l}\xi_i$ tende a minimizar as variáveis de folga ξ_i. Logo, o controle da relação de compromisso (“*trade-off”)* entre a complexidade da função de decisão e a quantidade de erros pode ser efetuada escolhendo-se um valor apropriado para a constante C (CORTES; VAPNIK, 1995).

Em outras palavras, atribuir valores grandes para C significa dar maior relevância ao número de erros e menor peso à margem do hiperplano. Nessa condição, poucos erros são permitidos e uma menor margem de classificação é gerada. Em contrapartida, quando se utilizam valores pequenos de C a situação inversa ocorre.

Resumidamente, a constante de regularização C caracteriza quão importantes serão as variáveis de folga, tornando o modelo mais ou menos sensível à presença de pontos mal classificados. Ressalta-se, no entanto, que seu valor deve ser positivo e especificado *a priori*.

De maneira similar ao algoritmo de margens rígidas, o modelo (41) também é transformado em seu correspondente dual. Utilizando procedimentos análogos aos descritos na subseção anterior, a formulação dual do algoritmo de classificação de margens flexíveis é dada por

$$\begin{aligned} &maximizar \quad \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &sujeito\ a: \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \\ &\qquad\qquad 0 \le \alpha_i \le \mathrm{C}, \qquad i = 1, \ldots, l. \end{aligned} \tag{42}$$

Para (42), das condições de KKT tem-se que

$$\alpha_i[y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) - 1 + \xi_i] = 0, \qquad i = 1, \ldots, l \tag{43}$$

$$\xi_i(\alpha_i - \mathrm{C}) = 0, \qquad i = 1, \ldots, l \tag{44}$$

Igualmente ao algoritmo de margens rígidas, os pontos de entrada para os quais $\alpha_i \neq 0$ são denominados vetores suporte. Entretanto, aqui faz-se necessário diferenciar os VS em duas categorias: os não limitados ($0 < \alpha_i < \mathrm{C}$) e os limitados, também chamados VS *Bound* ($\alpha_i = \mathrm{C}$).

Para a primeira categoria, obtém-se pela condição (44) que $\xi_i = 0$ e, consequentemente, de (43) decorre que esses VS estão situados sobre as margens de classificação (FIGURA 4). Já para o segundo caso, em que $\alpha_i = \mathrm{C}$, a equação (44) mostra que suas respectivas variáveis de folga são $\xi_i \neq 0$ e, dependendo do seu valor ($\xi_i > 1$ ou $0 < \xi_i \leq 1$), a localização do vetor, dada por (43), está do lado correto ou não da superfície de decisão (FIGURA 6). Logo, os VS *Bound* são considerados erros de classificação e, tal qual os VS não limitados, influenciam na construção do hiperplano separador. Novamente, a função de decisão $f(\boldsymbol{x})$ é expressa unicamente em termos dos VS e o vetor dos pesos $\boldsymbol{w}^*$, o hiperplano ótimo e o *bias* b^* são determinados pelas expressões (37) a (39).

2.3 CLASSIFICAÇÃO EM MÚLTIPLAS CLASSES

O SVC foi originalmente desenvolvido por Boser, Guyon e Vapnik (1992) para realizar a classificação binária. Contudo, existem duas abordagens que estendem a aplicação do SVC à classificação em múltiplas classes, são elas: a que defende construir e combinar diversos classificadores binários e a que opta por considerar todos os dados em um único problema de otimização. Tendo em vista que a formulação do SVC para múltiplas classes é proporcional ao número de classes, a última opção é normalmente a de maior custo computacional, pois trata-se da resolução de um problema de alta dimensão.

Este trabalho, no entanto, limita-se a apresentar o método “Um contra um”, do inglês “*One-against-one”*, de construção de vários classificadores binários, uma vez que esse será o procedimento empregado nesta tese. A escolha por essa técnica se dá unicamente pelos pacotes computacionais do LIBSVM - *Library for Support Vector Machines* (CHANG; LIN, 2011) aplicarem tal metodologia. Conforme será explicado no capítulo 4, todos os cálculos referentes ao SVC aqui realizados foram executados via LIBSVM. Entretanto, é interessante ressaltar que os desenvolvedores do LIBSVM

justificam o uso do “Um contra um” devido aos resultados de Hsu e Lin (2002) apontarem que esse é método mais adequado para as finalidades práticas.

Em Chaves (2006), Lima (2004) e Hsu e Lin (2002) encontram-se explicações pertinentes a outros métodos, que solucionam o SVC para múltiplas classes, como por exemplo: o “Um contra todos” (*“One-against-all”),* o DAGSVM (*Directed Acyclic Graph Support Vector Machine*) e o método de Crammer e Singer.

2.3.1 Método Um contra um

O método “Um contra um” foi introduzido por Knerr, Personnaz e Dreyfus (1990), mas suas primeiras aplicações voltadas ao algoritmo SVC foram dadas por Friedman (1996) e Krebel (1998). Nesse método, são construídos

$$\frac{k(k-1)}{2} \tag{45}$$

classificadores (SVC) binários, onde k é o número de classes. Para os dados de treinamento da i-ésima e j-ésima classes, com $i \neq j$, resolve-se o problema de classificação binária

$$\begin{aligned} &minimizar \quad \frac{1}{2}\left\|\boldsymbol{w^{ij}}\right\|^2 + \mathrm{C}\sum_{t=1}^{N}\xi_t^{ij} \\ sujeito\ a: &\left(\langle \boldsymbol{w^{ij}} \cdot \phi(\boldsymbol{x}_t)\rangle + b^{ij}\right) \geq 1 - \xi_t^{ij} \qquad se\ y_t = i \\ &\left(\langle \boldsymbol{w^{ij}} \cdot \phi(\boldsymbol{x}_t)\rangle + b^{ij}\right) \leq -1 + \xi_t^{ij} \quad se\ y_t = j \\ &\xi_t^{ij} \geq 0, \qquad t = 1, \ldots., N. \end{aligned} \tag{46}$$

Na prática, os vários modelos representados por (46), cuja quantidade é definida por (45), são solucionados por meio da sua formulação dual. Para combinar os resultados obtidos por todos os classificadores, utiliza-se a estratégia de voto sugerida por Friedman (1996). Nessa estratégia, também chamada de “*Max Wins”,* a classificação é realizada do seguinte modo: se o sinal de $\left(\langle \boldsymbol{w^{ij}} \cdot \phi(\boldsymbol{x}_t)\rangle + b^{ij}\right)$ for positivo, soma-se um voto para a classe i, caso contrário, para a classe j. Ao final dos

testes, o padrão $\boldsymbol{x}$ pertencerá à classe que recebeu o maior número de votos. Contudo, em casos de empate, Hsu e Lin (2002) sugerem que a classe de menor índice seja a selecionada[2].

2.4 A INFLUÊNCIA DA SELEÇÃO DE PARÂMETROS

A capacidade de generalização de um sistema de aprendizagem consiste na sua habilidade de prever adequadamente os padrões de teste. Essa aptidão está diretamente relacionada à forma como se executa a etapa de treinamento, da onde pode-se obter funções que melhor ou pior se ajustam as informações de entrada.

No caso do SVC, o desempenho do algoritmo está fortemente vinculado à seleção de seus parâmetros. Segundo Boser, Guyon e Vapnik (1992) e Cortes e Vapnik (1995), o uso de diferentes funções kernel possibilitam a construção de classificadores com distintas superfícies de decisão, cuja complexidade e quantidade de erros é determinada pelo valor da constante de regularização C.

Desta forma, a definição do hiperplano separador e a capacidade de generalização do SVC dependem significativamente da escolha de C, do tipo de kernel e de seus respectivos parâmetros. A seleção inadequada dessas variáveis pode levar a fenômenos indesejáveis como o *overfitting* e o *underfitting*, que afetam tanto a acurácia da classificação quanto o tempo do treinamento (WANG; HUANG; CHENG, 2014).

A ocorrência do *overfitting* consiste no ajuste excessivo do algoritmo aos dados de treinamento. Em outras palavras, encontra-se uma função de decisão muito complexa e flexível, adaptada inclusive aos ruídos dos padrões de entrada. Assim, por se especializar demasiadamente no conjunto de treinamento, o algoritmo não consegue prever corretamente as saídas dos dados de teste, perdendo sua capacidade de generalização. São características do *overfitting:* alta taxa de acertos na fase de treinamento e baixa acurácia na etapa de teste.

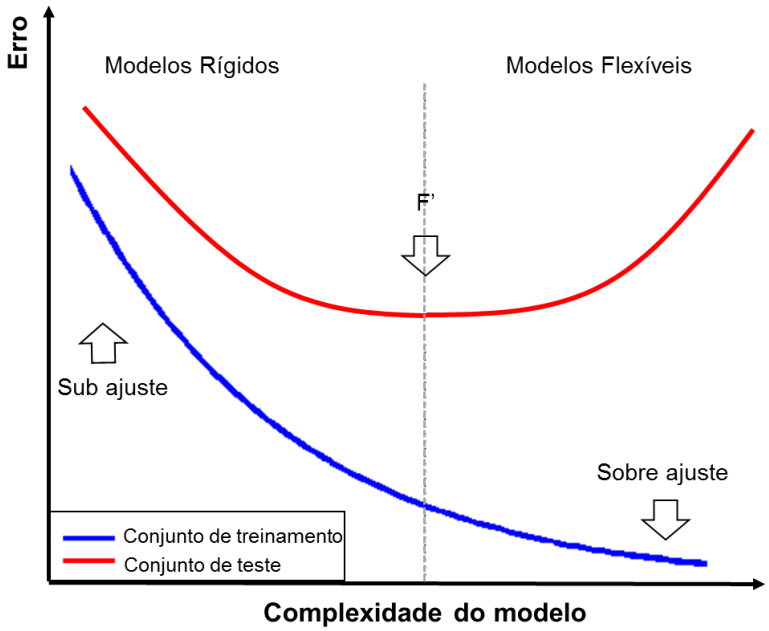
Em contrapartida, o *underfitting* acontece quando se determina uma função muito simples e rígida, cuja flexibilidade é insuficiente para capturar as informações importantes dos padrões de entrada. Desta forma, o modelo se ajusta muito pouco

[2] Nesta tese, cita-se a abordagem adotada por Hsu e Lin (2002) devido a Chih-Jen Lin ser um dos desenvolvedores do LIBSVM e ter se baseado nesse seu artigo de 2002 para definir alguns dos princípios de funcionamento do LIBSVM.

aos dados de treinamento, não os representando adequadamente. Assim, no *underfitting,* tanto a acurácia do treinamento quanto a de teste são baixas.

A FIGURA 7 ilustra os conceitos anteriormente explicados, exprimindo a relação entre a complexidade de um modelo e a sua taxa de erros.

FIGURA 7 - RELAÇÃO ENTRE A COMPLEXIDADE DE UM MODELO E A SUA TAXA DE ERROS



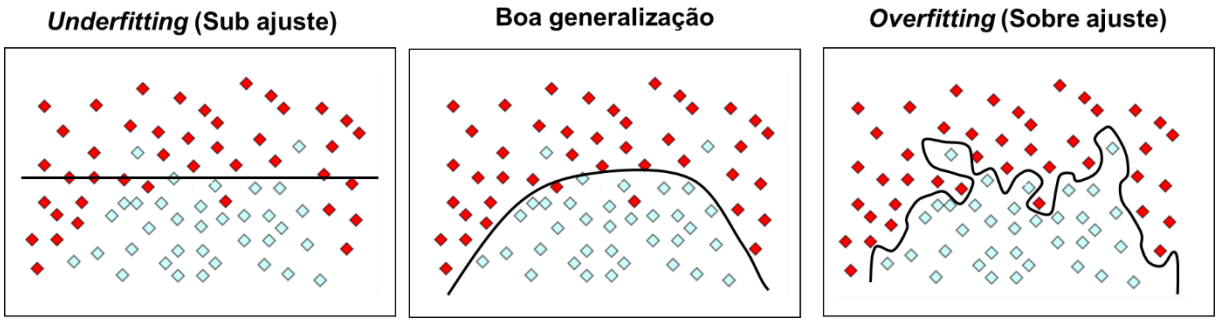
FONTE: LAVRENKO (2014).

Pela FIGURA 7, repara-se que as situações de *overfitting*, caracterizadas por funções muito flexíveis, ocorrem na extrema direta do gráfico onde a complexidade do modelo aumenta. Nessas ocasiões, os erros de treinamento são muito pequenos, próximos à zero, enquanto que os de teste são elevados. Já na extrema esquerda constam os casos de *underfitting,* dadas por funções rígidas e simples, que apresentam quantidades elevadas de erros tanto no treinamento quanto no teste. Por fim, no centro do gráfico verifica-se uma função F' que, apesar de não apresentar a menor taxa de erros possível no treinamento, é aquela que melhor generaliza os dados de teste.

Os conceitos da FIGURA 7 ficam ainda mais claros quando diretamente aplicados ao SVC. A FIGURA 8 compara um mesmo problema de classificação binária separado por diferentes hiperplanos. Na parte esquerda da FIGURA 8 consta um caso de *underfitting,* da onde percebe-se que o hiperplano encontrado é bastante simples e não representa adequadamente o conjunto de treinamento. Já na parte direita,

visualiza-se um hiperplano muito complexo e extremamente adaptado aos padrões de entrada, o que caracteriza uma típica situação de *overfitting.* Finalmente, na parte central da FIGURA 8 ilustra-se uma função de decisão de boa capacidade de generalização, que é sem dúvida a melhor opção entre as apresentadas.

FIGURA 8 - A GENERALIZAÇÃO DE DIFERENTES HIPERPLANOS SEPARADORES

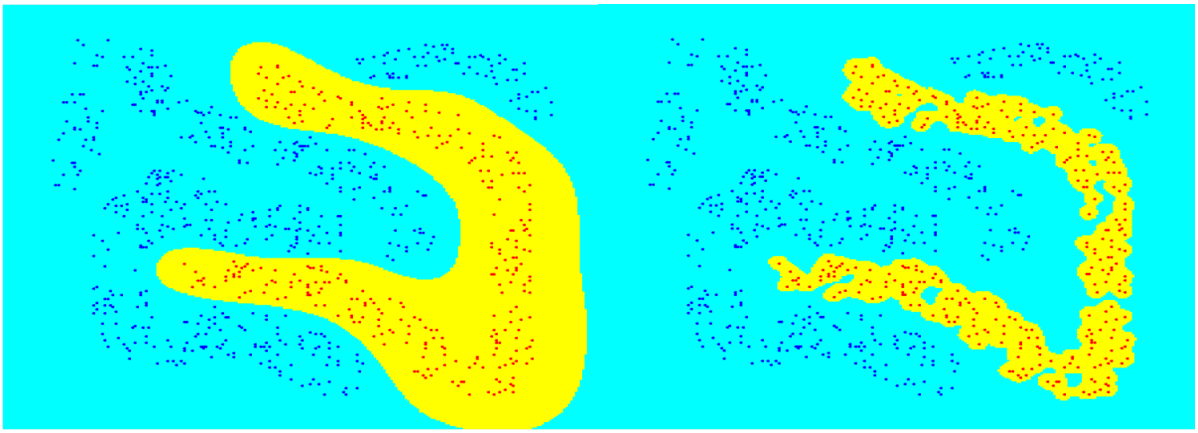


FONTE: A autora (2016).

No SVC, a ocorrência do *overfitting* é mais facilmente perceptível que a do *underfitting*, pois ela impacta diretamente na quantidade de vetores suporte do problema. Desta forma, o número de VS empregado na construção do hiperplano separador fornece informações importantes quanto a qualidade da função. De acordo com Boser, Guyon e Vapnik (1992), essa porção se relaciona com a efetiva capacidade do classificador, pois quantidades elevadas de vetores suporte, que se aproximam ou se igualam a totalidade do conjunto de treinamento, indicam um sobre ajuste do sistema (*overfitting)*, enquanto que um baixo número de VS reflete numa melhor generalização dos dados.

Para ilustrar essa relação, a FIGURA 9 mostra o impacto da escolha dos parâmetros na construção do hiperplano separador do conjunto *Fourclass* (HO; KLEINBERG, 1996), cuja complexidade é diretamente proporcional ao número de vetores suporte encontrado. Nesse exemplo, empregou-se para ambos os casos (FIGURA 9 à esquerda e à direita) a função kernel gaussiano. Os valores de γ e da constante C foram arbitrariamente determinados com o intuito de evidenciar a influência da má seleção de parâmetros (sem critérios) na qualidade da função.

FIGURA 9 - SEPARAÇÃO ADEQUADA VERSUS *OVERFITTING*



FONTE: BELTRAMI e SILVA (2015).

O treinamento do SVC da FIGURA 9 à esquerda foi realizado com os parâmetros C= 100 e γ= 10 e o da direita com C= γ= 1000. Para o primeiro caso, encontrou-se um hiperplano separador de melhor generalização determinado por 43 vetores suporte, enquanto que para o segundo, da onde nota-se uma superfície demasiadamente especializada, resultaram 799 VS. Considerando que o conjunto de treinamento era composto por 862 padrões, as respectivas funções de decisão foram construídas com 4,99% e 92,69% dos dados, onde o último caracteriza a ocorrência de *overfitting*. Na FIGURA 9, os pontos azuis representam os dados da classe -1 e os vermelhos os da classe +1.

Apesar da FIGURA 9 exemplificar a influência de ambos os parâmetros, constante de regularização C e γ do kernel gaussiano, na performance do SVC, Lin *et al* (2008a, 2008b) destacam que γ é o que mais impacta nos resultados da classificação, pois seu valor se relaciona diretamente com a partição do espaço de características.

2.4.1 A preferência pelo kernel gaussiano e a influência do seu parâmetro no comportamento assintótico do SVC

A função kernel gaussiano (RBF), em virtude do seu amplo domínio de convergência e vasta gama de aplicações (PANG *et al*.,2011; WANG *et al*.,2014), é a mais empregada no SVC. Por essa razão, quando se trata da seleção de parâmetros do SVC, a preocupação dos usuários limita-se a determinar valores adequados para

a constante de regularização C e para o parâmetro[3] γ ou σ^2 do kernel gaussiano. Dificilmente questiona-se qual é o melhor kernel a adotar, apenas parte-se do princípio que o gaussiano será o escolhido. Hsu, Chang e Lin (2010), com base nos estudos de Vapnik (1995), Keerthi e Lin (2003) e Lin e Lin (2003), explicam motivos que justificam essa preferência:

1. O kernel gaussiano, diferentemente do linear, pode lidar com dados não linearmente separáveis. Ainda, em certas condições, o kernel linear é considerado um caso especial do RBF (KEERTHI; LIN, 2003).
2. O kernel sigmoidal para alguns valores de parâmetros comporta-se como o RBF (LIN; LIN, 2003) e para outros ele não é válido, pois não consiste em um produto interno de dois vetores (VAPNIK, 1995).
3. Visto que o número de parâmetros internos ao kernel influencia a complexidade da seleção de modelos, prefere-se o gaussiano ao polinomial não homogêneo por ele ter menos parâmetros a definir.
4. Por fim, o kernel polinomial apresenta maiores dificuldades numéricas que o RBF, já que seu valor pode tender a infinito ou a zero quando seu grau é muito alto.

Explicitadas as vantagens do kernel gaussiano frente aos demais, torna-se necessário compreender como as escolhas do seu parâmetro (γ ou σ^2) e da constante C impactam no desempenho do SVC. Nesse contexto, Keerthi e Lin (2003) deram grande contribuição ao tema, pois analisaram o comportamento assintótico do SVC para os casos em que C e σ^2 assumem valores muito grandes ou bem pequenos. A partir desse estudo, Keerthi e Lin (2003) evidenciaram que:

- Severo *underfitting* ocorre (isto é, todo o conjunto de dados é atribuído à classe majoritária[4]) quando: a) σ^2 é fixo e $C \to 0$; b) $\sigma^2 \to 0$ e C é fixado em um valor suficientemente pequeno; c) $\sigma^2 \to \infty$ e C é fixo.
- Grave *overfitting* ocorre (ou seja, somente pequenas regiões em torno dos dados da classe minoritária são consideradas pertencentes a ela, enquanto que o restante do espaço é classificado como classe majoritária) quando: $\sigma^2 \to 0$ e C é fixado em um valor suficientemente grande.

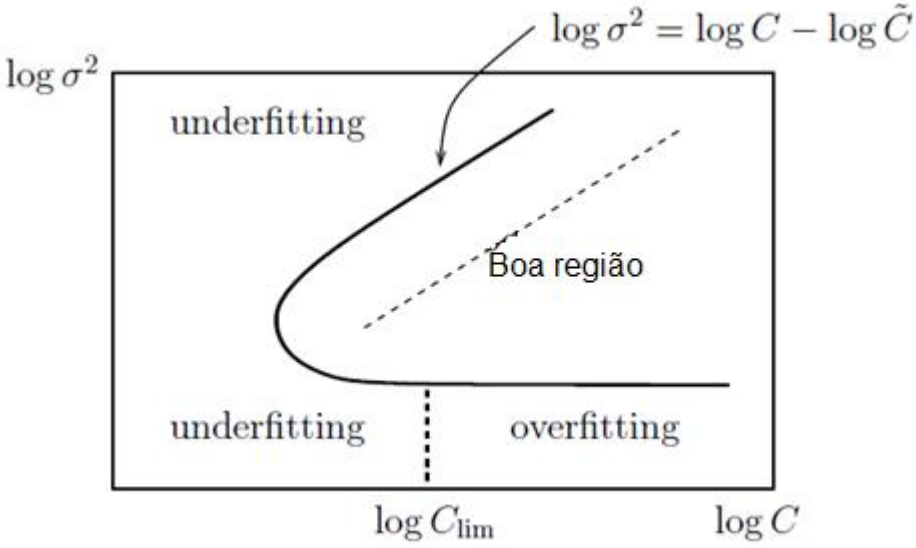
[3] Ver equações destacadas no QUADRO 1.

[4] Classe majoritária é aquela que possui o maior número de dados de treinamento pertencentes a ela. É importante destacar que a pesquisa parte do princípio que as duas classes consideradas nunca possuem a mesma quantidade de padrões.

- Se σ^2 é fixo e $C \to \infty$ o SVC separa o conjunto de treinamento estritamente em duas classes. Considera-se isso um caso de *overfitting* se os dados analisados tiverem ruído.
- Se $\sigma^2 \to \infty$ e $C = \tilde{C}\sigma^2$ onde $\tilde{C}$ é fixo, então o SVC com kernel RBF converge para o SVC linear de parâmetro $\tilde{C}$.

As constatações feitas por Keerthi e Lin (2003) possibilitaram o entendimento e a caracterização do espaço de busca dos parâmetros C e σ^2. A reunião dessas observações mostrou a existência de uma curva de erro de generalização, que separa a área de busca de C e σ^2 em duas regiões: uma definida pelo *underfitting e* pelo *overfitting* e outra denominada boa região. É nessa última que se encontra o par de parâmetros ótimo, que propicia a construção do hiperplano separador de melhor capacidade de generalização. A FIGURA 10 ilustra esse espaço e suas referidas regiões.

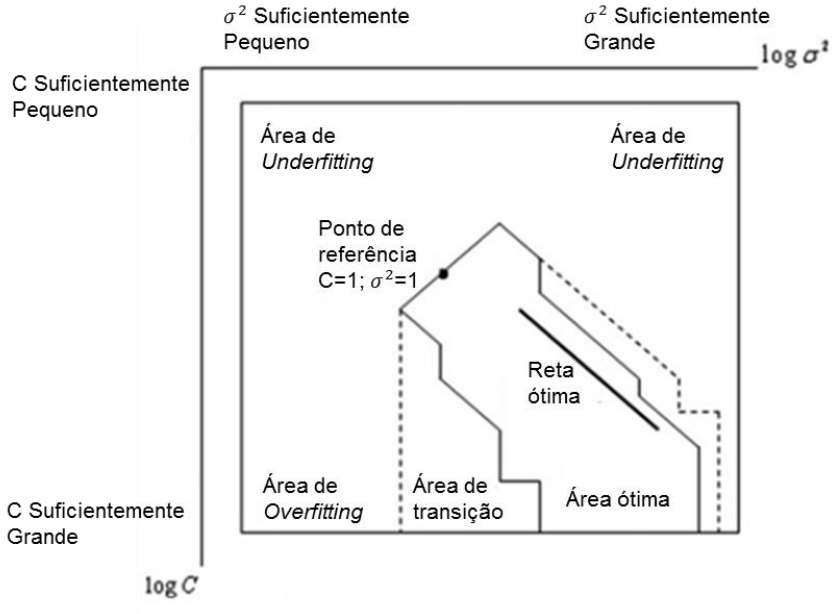
FIGURA 10 - ESPAÇO DE BUSCA DOS PARÂMETROS (C, σ^2) E A EXISTÊNCIA DA BOA REGIÃO



FONTE: Adaptado de KEERTHI e LIN (2003).

Zhao *et al.* (2012), motivados pelo trabalho de Keerthi e Lin (2003), deram continuidade ao estudo da influência dos parâmetros (C, σ^2) no comportamento do SVC e constataram novas propriedades do seu espaço de busca. Desta forma, as áreas descritas na FIGURA 10 foram ainda mais detalhadas, segundo mostra a FIGURA 11.

FIGURA 11 - DISTRIBUIÇÃO DAS ÁREAS DO PLANO C – σ^2



FONTE: Adaptado de ZHAO *et al.* (2012).

Pela FIGURA 11, cuja rotação é de 270º em relação à FIGURA 10, constata-se que, além da boa região e das zonas de *underfitting* e *overfitting,* o plano C – σ^2 é constituído por uma área de transição e um ponto de referência, situado na fronteira da região ótima, cuja coordenada é (1, 1). Tal detalhamento foi obtido a partir da avaliação de diversos conjuntos de dados dos repositórios UCI *Machine Learning* (LICHMAN, 2013) e IDA *Benchmarch* (RÄTSCH, 1999).

Devido ao embasamento teórico de Keerthi e Lin (2003), que se fundamenta em inúmeros teoremas, o modelo de seleção de parâmetros desenvolvido nesta tese tem como princípio norteador a existência das regiões destacadas na FIGURA 10. Em virtude do modo de funcionamento da técnica *quadtree,* que será explicado nos capítulos 3 e 4*,* o modelo aqui proposto não necessita de todos os detalhes da FIGURA 11 e, portanto, baseia-se somente na FIGURA 10.

2.5 MODELOS DE SELEÇÃO DE PARÂMETROS

A tarefa de procurar parâmetros ótimos para o SVC, em relação a suas medidas de desempenho, é denominada seleção de modelos do SVC, do inglês, SVC *model selection problem* (CHAPELLE; VAPNIK, 1999). Ao estudar esse problema, o desafio está em encontrar maneiras simples de executar tal atividade em tempos computacionalmente aceitáveis. Nesse sentido, pesquisadores têm se

dedicado ao desenvolvimento de diferentes métodos de determinação de parâmetros do SVC.

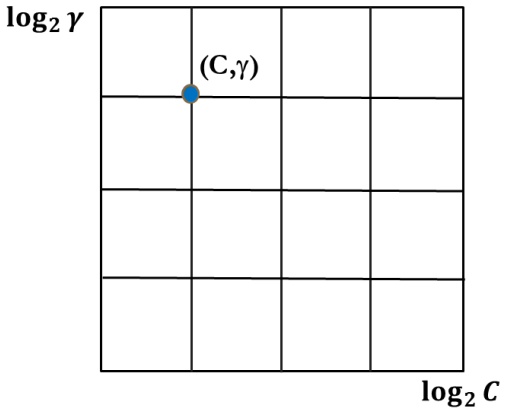
De acordo com Kapp, Sabourin e Maupin (2012), esses modelos diferem basicamente em dois aspectos: critério de seleção e metodologia de busca adotados. Entende-se por critério de seleção a medida de avaliação de erro de generalização empregada para guiar o processo de procura de parâmetros, como por exemplo: contagem de vetores suporte (VAPNIK, 1995), *radius margin bound* (VAPNIK, 1998), *span bound* (CHAPELLE; VAPNIK, 1999) e taxa de validação cruzada (HSU; CHANG; LIN, 2010). Já a metodologia de busca refere-se a técnica aplicada para encontrar os referidos parâmetros, como: descida de gradiente (CHAPELLE *et al.*, 2002), busca por *grid*, metaheurísticas em geral e etc.

Dentre as técnicas existentes, a busca por *grid* (BG), também conhecida como *grid search,* é a mais utilizada. A popularidade da BG entre seus usuários se dá pela sua facilidade de uso e pelo fornecimento de bons resultados (DEVOS *et al*, 2009). Contudo, por analisar todas as combinações de parâmetros em seu espaço discreto de busca, a mesma apresenta alto custo computacional, o que a torna inviável para a avaliação de grandes conjuntos de dados. Mas, apesar das suas limitações, a BG apresenta grande relevância na área de seleção de modelos, pois ela serve de referência para a validação de novas metodologias, que ao serem propostas comparam os seus resultados com os da BG. Diante disso, na próxima subseção explicam-se as características da BG e, na sequência, apresentam-se os demais métodos.

2.5.1 Busca por *grid*

A busca por *grid* (BG) tem por objetivo encontrar o par de parâmetros ótimos do SVC, constante de regularização C e γ do kernel gaussiano, em uma malha ou grade, conforme ilustra a FIGURA 12. Para aprimorar a habilidade de generalização do SVC, a BG emprega em seu processo de investigação a metodologia de validação cruzada (VC) do tipo *k-fold.* Segundo Hsu, Chang e Lin (2010), esse procedimento visa prevenir a ocorrência de *overfitting.*

FIGURA 12 - REPRESENTAÇÃO DA BUSCA POR *GRID*



FONTE: A autora (2016).

Assim, para cada ponto (C, γ) do espaço de busca, a BG efetua uma operação de validação cruzada. Basicamente, a VC *k-fold* consiste em dividir o conjunto de treinamento em *k* subconjuntos de mesmo tamanho, treinar o algoritmo com *k* – 1 subconjuntos e testá-lo com o remanescente. Esse processo é repetido *k* vezes de forma a garantir que todos os subconjuntos sejam avaliados. Ao final das *k* iterações, calcula-se a precisão do modelo, contabilizando-se o percentual de acertos. De acordo com Kapp, Sabourin e Maupin (2012), o valor *k* = 5 é o mais utilizado, porém verifica-se que muitos autores trabalham com *k* = 10. Em resumo, a busca por *grid* é descrita pelos passos do QUADRO 2.

QUADRO 2 - PSEUDOCÓDIGO DA BUSCA POR *GRID*

MÉTODO: BUSCA POR *GRID*
1. Considerar uma malha discreta (*grid*) no espaço de coordenadas $(C' = \log_2 C, \gamma' = \log_2 \gamma)$.
2. Para cada par de parâmetros (C, γ) do espaço de busca, realizar uma validação cruzada *k-fold* no conjunto de treinamento.
3. Escolher o par (C, γ) que resulte na maior taxa de acertos de validação cruzada.
4. Usar os parâmetros do passo 3 para criar o modelo SVC.

FONTE: Adaptado de AKAY (2009).

Em virtude da BG ser uma busca exaustiva, o tamanho do seu espaço de busca, determinado pela faixa de procura de parâmetros e pela espessura do *grid,* influencia diretamente no tempo computacional despendido pela técnica (LI *et al,* 2012). Ressalta-se ainda que a dimensão do conjunto de treinamento (número de dados e de características) e o valor *k* da VC *k-fold* também impactam nesse custo*,* pois quanto maior os seus valores, maior será a quantidade de cálculos efetuados.

5.2.2 Demais métodos

Nesta seção, destacam-se os demais métodos de seleção de parâmetros do SVC. Devido às várias opções existentes, apresentam-se os mais notáveis e os mais recentemente desenvolvidos. Conforme será visto, muitos são os desafios encontrados nessa área de estudo. Logo, não há modelos perfeitos e, portanto, todos possuem suas vantagens e desvantagens.

Uma das primeiras técnicas evidenciadas na literatura foi a de Chapelle *et al*. (2002), que consiste num método automático de determinação de parâmetros, que minimiza estimativas de erro de generalização por meio de um algoritmo de descida do gradiente. Apesar dessa metodologia reduzir significativamente a complexidade computacional, ela exige uma função objetivo diferenciável em relação aos parâmetros do SVC e normalmente se depara com ótimos locais.

Keerthi e Lin (2003), após estudar o comportamento assintótico do SVC e desenhar a FIGURA 10, desenvolveram uma heurística para encontrar (C, σ^2) de maneira mais rápida que a busca por *grid*. O benefício desse método, conhecido por busca linear, está em testar cerca de $2r$ pontos enquanto que o *grid* uniforme, de tamanho $r \times r$, avalia r^2 combinações. Nessa técnica, trata-se de construir uma reta de inclinação unitária, que corta a boa região na sua parte central (linha tracejada da FIGURA 10), e procurar sobre ela o par de pontos ótimos. Para isso, executa-se os passos descritos no QUADRO 3.

QUADRO 3 - PSEUDOCÓDIGO DA BUSCA LINEAR

MÉTODO: BUSCA LINEAR
1. Procurar o melhor valor da constante de regularização C do SVC com kernel linear e chamá-la de $\tilde{C}$.
2. Fixar o valor de $\tilde{C}$ do passo 1 e buscar o melhor par (C,σ^2) satisfazendo $log\,\sigma^2 = \log C - log\ \tilde{C}$, usando o kernel gaussiano.

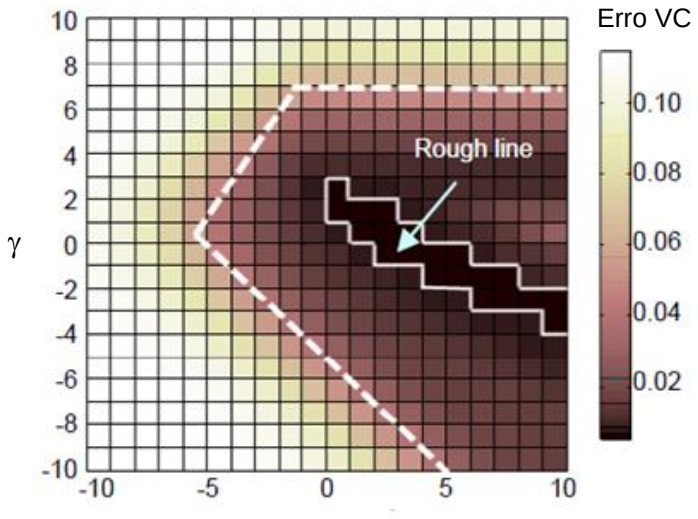
FONTE: KEERYHI e LIN (2003).

A metodologia de Keerthi e Lin (2003) fundamenta-se no fato de que se $\sigma^2 \rightarrow \infty$, o SVC com kernel gaussiano comporta-se como o SVC linear. Para avaliar o desempenho da heurística, os autores empregaram a VC 5-*fold* e compararam seus resultados com o da BG. Observou-se que o procedimento proposto é competitivo ao *grid* no que diz respeito à acurácia, pois para cinco dos dez conjuntos binários

avaliados obtiveram-se as mesmas taxas de erros. Nos demais, a BG apresentou superioridade em três conjuntos (4,22%, 10,07% e 25,96%) e a busca linear nos dois restantes (4,62% e 6,00%). Entretanto, a vantagem do método de Keerthi e Lin (2003) se deu no menor tempo despendido, pois enquanto a BG analisou 441 pares de parâmetros a sua heurística testou apenas 54.

Pang *et al.* (2011) fundamentando-se em Keerthi e Lin (2003) elaboraram um método para determinar os parâmetros (C, γ) denominado nova busca linear. O termo nova refere-se as ressalvas e adaptações feitas acerca de Keerthi e Lin (2003). Tais restrições surgiram à medida que Pang *et al.* (2011), após analisarem vários conjuntos de dados e construírem seus respectivos gráficos nos moldes da FIGURA 10, concluíram que os parâmetros ótimos não estavam necessariamente localizados em uma reta de inclinação unitária, mas sim em uma área mais larga, ilustrada na FIGURA 13. É nessa região, intitulada pelos autores de "*rough line* " em que estão situados os pontos de menores erros de validação cruzada.

FIGURA 13 - "*ROUGH LINE*" CORTANDO A BOA REGIÃO



FONTE: PANG *et al*. (2011).

Assim, Pang *et al.* (2011) consideram que o método de Keerthi e Lin (2003) pode em alguns casos identificar a solução incorreta, visto que eles se limitam a procurá-la ao longo de uma reta. Desta forma, os autores propõem, por meio da marcação de pontos, delimitar a região da "*rough line*", que irradia da parte central da boa região para pontos adjacentemente posicionados abaixo e à direita, e na sequência encontrar o par (C, γ) na região demarcada. Para avaliar a sua metodologia,

Pang *et al.* (2011) a compararam com a BG e com a busca linear de Keerthi e Lin (2003), utilizando dados do repositório UCI e IDA *Benchmark.* Concluiu-se que o método sugerido, além de localizar corretamente a área de parâmetros ótimos, diferentemente de Keerthi e Lin (2003), é capaz de encontrar soluções tão boas quanto as da busca por *grid*, mas em tempos muito menores. Para cinco dos oito conjuntos avaliados, a BG e a nova busca linear encontraram as mesmas taxas de validação cruzada. Nos outros três, o desempenho da BG foi de 1,77%, 4,69% e 6,09% superior. Contudo, em todos os conjuntos, a BG levou mais tempo para encontrar a solução, demorando em torno de 2 a 14 vezes mais que a técnica proposta. Em relação a Keerthi e Lin (2003), somente em um dos conjuntos Pang *et al.* (2011) e ela determinaram a mesma solução (C, γ). Isso, além de evidenciar diferentes localizações de parâmetros por parte da *"rough line"* e da reta de Keerthi e Lin (2003), resultou numa diferença média de acurácia de 9,02%.

Diversos são os métodos de determinação de parâmetros do SVC baseados em metaheurísticas, como por exemplo: algoritmo genético, nuvem de partículas, colônia de formigas, dentre outros. Todavia, autores como Pang *et al* (2011) e Wang, Huang e Cheng (2014) criticam o uso dessas metodologias devido aos seus desempenhos também dependerem da correta seleção de seus parâmetros, como por exemplo: tamanho da população inicial, taxas de *crossover* e de mutação, para o algoritmo genético. Contudo, conforme se verá adiante, se bem ajustados os parâmetros das metaheurísticas, essas se mostram bastante eficientes na solução do problema de seleção de modelos do SVC.

Lin *et al.* (2008a) criaram um método usando a técnica *simulated annealing* (SA), chamado SA-SVM, que determina simultaneamente as principais características dos dados de entrada e os parâmetros (C, γ) do SVC. No SA-SVM, a finalidade da seleção de características (atributos) é identificar e eliminar componentes redundantes do vetor de entrada, de forma a tornar o processo de classificação mais rápido e preciso. Isso é possível já que algumas características da amostra não agregam informações adicionais e contém um alto nível de ruído. Os autores compararam o desempenho do SA-SVM com o da busca por *grid*, avaliando onze conjuntos de dados do repositório UCI *Machine Learning.* Como a solução do SA-SVM não é determinística, para cada conjunto considerado, o modelo foi executado cinco vezes a fim de se obter uma taxa média de acerto. Além disso, o SA-SVM foi testado com a seleção de características e sem a mesma, de maneira a analisar os

benefícios desse procedimento. A partir dos resultados, concluiu-se que a seleção de características proporcionou uma maior eficiência ao SVC, aumentando em média 3,5% da sua acurácia. Entretanto, o uso dessa técnica dobrou o tempo computacional do SA-SVM, visto que foram necessárias mais iterações para explorar o espaço de busca de soluções. Verificou-se também que a performance do método proposto superou o da BG em todos os casos avaliados. Em média, a acurácia do SA-SVM com e sem seleção de características foi respectivamente 8,90% e 5,20% maior que a da busca por *grid.* Em relação a essa última, não constam informações a respeito do seu tempo de processamento.

Huang e Wang (2006) propuseram um método que seleciona simultaneamente as características dos vetores de treinamento e os parâmetros (C, γ) do SVC usando o algoritmo genético, do inglês, *Genetic Algorithm* (GA). Para validar o método proposto, os autores avaliaram onze conjuntos de dados do repositório UCI e compararam seus resultados com os da busca por *grid.* Os experimentos do GA foram realizados via *Matlab* e LIBSVM e os da BG via *Phyton.* Para todos os conjuntos estudados, a acurácia do GA foi em média 9,02% superior à da BG. Exclusivamente para os conjuntos de classificação binária, analisou-se o poder de discriminação de ambas as técnicas. Em quatro dos sete conjuntos binários, o algoritmo genético apresentou maior taxa de acerto tanto para as classes positivas quanto negativas. Contudo, apesar de não apresentarem os valores de tempo despendido, os autores destacaram que o tempo médio de execução do método proposto foi somente um pouco menor que o da BG. Huang e Wang (2006) justificam que essa diferença não foi expressiva devido ao uso do *Matlab,* que normalmente é mais demorado do que os outros sistemas, neste caso o *Phyton*.

Zhao *et al.* (2011) integraram os conceitos do comportamento assintótico do SVC de Keerthi e Lin (2003) ao GA e elaboraram um método denominado algoritmo genético com cromossomos de atributos[5] (*feature chromosomes).* O objetivo dessa metodologia é, além de determinar as principais características dos dados de entrada, direcionar a procura dos parâmetros (C, γ) sobre a reta de inclinação unitária proposta por Keerthi e Lin (2003), ilustrada na FIGURA 10. Para isso, os autores introduziram a relação $log\sigma^2 = \log C - log\ \tilde{C}$ ao GA para que os seus cromossomos pudessem

[5] Zhao *et al.* (2011) explicam que a diferença fundamental entre o seu método e a de um algoritmo genético comum é a aplicação dos conceitos de Keerthi e Lin (2003).

selecionar os parâmetros desejados. O modelo, cujo desenvolvimento se deu no ambiente *Matlab* juntamente com o LIBSVM, teve seu desempenho comparado com os métodos de Keerthi e Lin (2003), Huang e Wang (2006) e busca por *grid*. Utilizando-se conjuntos de dados do repositório UCI e IDA *Benchmark*, evidenciou-se que o método proposto foi superior às três técnicas comparadas. Em média, a sua acurácia foi respectivamente 1,03%, 1,83% e 7,67% maior. Ainda, quando confrontado a performance do SVC com e sem seleção de características, constatou-se que a primeira aumentou, em média, 3,26% da sua acurácia. Em relação ao tempo computacional, os autores apenas o compararam com a da BG, apresentando apenas os valores de quatro dos onze conjuntos avaliados. Para três desses quatro, o método proposto foi em média 30,13% mais rápido que a BG.

Zhao *et al.* (2012), após detalharem o espaço de busca dos parâmetros (C, σ^2) conforme ilustra a FIGURA 11, propuseram um método de seleção de parâmetros baseado em algoritmos genéticos com mudança de área. Desejando-se evitar as zonas de *underfitting* e *overfitting,* o método considera o ponto de referência (1,1) como a extremidade inicial da área de busca. Nesse método, o GA procura parâmetros da parte superior a esquerda para a inferior a direita, seguindo a direção da reta de inclinação unitária, da onde deseja-se obter os parâmetros ótimos. Assim, diminui-se a área de procura a cada iteração. Os experimentos foram realizados por meio do *Matlab* em conjunto com o LIBSVM e os resultados obtidos foram comparados com os da busca linear de Keerthi e Lin (2003). Para os sete conjuntos do repositório IDA *Benchmark* avaliados, o método apresentou em média 8,55% a menos de erros na predição dos conjuntos de teste. No artigo não constam informações quanto ao tempo de execução das técnicas.

Zhang, Chen e He (2010), empregando o algoritmo de colônia de formigas, do inglês *Ant Colony Optimization* (ACO), desenvolveram um método denominado ACO-SVM, que determina os parâmetros (C, σ) do SVC. Com essa técnica, deseja-se que as formigas produzam soluções (C, σ) de baixos erros de classificação e depositem sobre elas quantidades de ferormônio proporcionais a sua qualidade, fazendo com que os melhores parâmetros sejam os mais atrativos para as próximas formigas. O ACO-SVM parte de um *grid* pré-definido, onde cada ponto da malha representa um par (C, σ). A cada iteração, reduz-se o espaço de busca (tamanho do *grid*) à vizinhança do melhor ponto, limitando a quantidade de parâmetros a serem testados pelo ACO-SVM. Para avaliar o desempenho da técnica proposta, os autores

analisaram cinco conjuntos binários, retirados dos repositórios UCI e IDA *Benchmark,* e compararam seus resultados com os da busca por *grid.* Em dois dos cinco conjuntos avaliados, o ACO-SVM e a BG obtiveram o mesmo percentual de erro de classificação. Já nos demais, o ACO-SVM apresentou desempenho superior ao da BG, pois as suas taxas de erros foram de 1,41%, 15,79% e 33,25% menores. Ainda, o ACO-SVM resolveu todos os problemas de forma mais rápida, reduzindo cerca de 5,22% a 64,08% do tempo de execução.

Zhang *et al*, (2015), em continuidade às pesquisas de Zhang, Chen e He (2010), elaboraram um método de colônia formigas, que seleciona simultaneamente as características dos dados de entrada e os parâmetros (C, σ) do SVC. O algoritmo foi aplicado para diagnosticar falhas em máquinas rotativas, sendo essas: um sistema de rotor e rolamentos de rolos de uma locomotiva. Para essas respectivas aplicações, o método proposto classificou corretamente, em média, 100% e 95,83% dos defeitos avaliados.

Lin *et al.* (2008b), Huang e Dun (2008), Kapp, Sauborin e Maupin (2012) e Jiang e Zou (2013) desenvolveram diferentes modelos de seleção de parâmetros do SVC empregando a técnica nuvem de partículas, do inglês, *Particle Swarm Optimization* (PSO). O conceito do PSO é simular o comportamento social dos pássaros, onde cada indivíduo (partícula) muda a sua direção baseando-se na sua própria experiência e na do restante do grupo.

Dentre os quatro métodos de PSO anteriormente citados, o de Lin *et al.* (2008b) é o mais simples. Esses autores aplicaram a técnica PSO, destinada à problemas contínuos, para selecionar as principais características dos dados de entrada e determinar os parâmetros (C, γ) do SVC. No que se refere à seleção de características, como a saída do PSO foi considerada contínua, se o valor da variável fosse maior que 0,5 escolhia-se a característica e caso contrário não. Para avaliar o desempenho do método proposto, denominado PSO+SVM, diversos conjuntos do repositório UCI foram analisados e seus resultados foram comparados com os métodos: busca por *grid* e modelo de algoritmo genético de Huang e Wang (2006). Concluiu-se que, adotando-se a seleção de características, a performance do PSO+SVM foi similar à de Huang e Wang (2006), pois para seis dos dez conjuntos avaliados o PSO foi superior ao algoritmo genético e para os demais quatro a situação inversa ocorreu. Além disso, a variação média entre as acurácias do PSO e do AG não foi superior à 1%. Já em relação a busca por *grid*, avaliaram-se dezessete

conjuntos e em todos eles o PSO+SVM, com e sem seleção de características, superou a BG. Para a primeira e segunda condição, as variações médias entre as acurácias do PSO e da BG foram respectivamente 8,39% e 5,21%, mostrando novamente uma melhoria nos resultados do SVC proporcionada pela seleção de características.

Huang e Dun (2008) combinaram as versões discretas e contínuas do PSO para determinar as características mais relevantes dos vetores de entrada e o par de parâmetros (C, γ) do SVC. Nesse trabalho, empregou-se um conjunto de dados criado artificialmente para verificar o impacto de três diferentes estratégias de definição de peso inercial (constante, decrescente e randômica) na acurácia do algoritmo proposto, intitulado PSO-SVM híbrido. Dentre as trinta características desse conjunto artificial, apenas cinco eram consideradas importantes para classificação, as demais constituíam-se de ruídos. Uma vez que o PSO-SVM híbrido foi capaz de selecionar corretamente as cinco características e não houve diferenças estatisticamente significativas entre as acurácias do SVC, quando aplicadas as três estratégias de peso inercial, decidiu-se por adotar na configuração final do PSO um peso inercial constante. A partir dessa definição, os autores implementaram o PSO-SVM híbrido via sistema de processamento distribuído (arquitetura paralela) a fim de reduzir o tempo computacional da técnica. Para tal, foram empregadas linguagens de programação, como: Matlab, XML (*eXtensible Markup Language*), SOAP (*Simple Object Acess Protocol*), e pacotes computacionais do LIBSVM. Para avaliar a proposta, os autores estudaram o conjunto *German* do UCI e compararam seus resultados com o método de algoritmo genético de Huang e Wang (2006). A acurácia do PSO-SVM híbrido foi inferior ao do GA, sendo 77,82 ± 3,98% contra 85,6% do GA. Apesar dos tempos computacionais de ambas técnicas não terem sido confrontados, os autores destacaram que a rapidez do seu método depende de fatores como: servidor de banco de dados, número de clientes agentes, capacidade da rede, dentre outros.

Jiang e Zou (2013) integraram as técnicas PSO e *simulated annealing* (SA) para determinar os parâmetros (C, σ) do SVC. Nesse método, todos os cálculos referentes à avaliação dos parâmetros e melhoria da solução foram efetuados pelo PSO. O SA foi utilizado apenas para evitar a convergência do PSO para mínimos locais. Uma vez que o objetivo da proposta era reconhecer padrões na área médica, os autores adaptaram a função *fitness* do PSO para fornecer resultados em termos

da fração de diagnósticos verdadeiros positivos e de falsos positivos. Para avaliar o desempenho do método, analisaram-se dois conjuntos de dados retirados do LIBSVM (*Diabetes* e *Heart Disease*) e imagens abdominais de pacientes chineses, que apresentavam condição de normalidade, câncer e cisto no fígado. Ao comparar os resultados da metodologia proposta com os de um método PSO tradicional, verificou-se que a acurácia da primeira foi superior à do segundo em três dos quatro conjuntos avaliados. No entanto, evidenciou-se que as variações entre as acurácias não foram superiores à 1%, com exceção do conjunto: câncer no fígado *versus* condição normal, que o modelo de PSO + SA superou o de PSO em 35,32%.

Em sua pesquisa, Kapp, Sabourin e Maupin (2012) consideraram a seleção de parâmetros (C, γ) do SVC como um problema dinâmico de otimização. Nessa concepção, entende-se que parâmetros anteriormente determinados necessitam ser revistos ou recalculados à medida que se deseja apresentar novos padrões de treinamento ao algoritmo. Os autores destacam que esse tipo de modelo é bastante válido para análises de séries temporais, em que o sistema deve ser capaz de se adaptar a novas entradas de informações, bem como o reconhecimento de assinaturas e diagnósticos de câncer. Desta forma, Kapp, Sabourin e Maupin (2012) elaboraram um modelo dinâmico de seleção de parâmetros composto por um detector de mudança, um *grid* modificado e uma nuvem de partícula dinâmica (DPSO), que é uma adaptação do PSO tradicional. Assim, para a primeira amostra de dados, determina-se o par (C, γ) via DPSO e, no caso de futuras entradas, utiliza-se o módulo detector para verificar se o par (C, γ) corrente continua viável ou não à aplicação. O objetivo desse módulo é analisar se a manutenção dos parâmetros altera significativamente o desempenho do SVC, evitando procuras desnecessárias e gasto de tempo computacional. Caso necessite-se de uma nova busca de parâmetros, essa é realizada via *grid* adaptado, procurando-se pares (C, γ) somente no entorno da antiga solução. Julgando-se necessário, a nova solução pode ser ainda melhorada pela DPSO.

Para avaliar o desempenho do método proposto, Kapp, Sabourin e Maupin (2012) estudaram bases de dados do Projeto Statlog (MICHIE; SPIEGELHALTER, TAYLOR, 1994), do repositório UCI e o conjunto *Circle and Square* de Carpenter, Grossberg e Reynolds (1991). Com o intuito de acelerar o tempo de execução dos experimentos realizados, a metodologia sugerida foi implementada em uma

arquitetura de processamento paralelo. Compararam-se os resultados obtidos com as das seguintes estratégias: tradicional busca por *grid* (BG), busca por *grid* 1*st* (1-*st* BG), nuvem de partículas clássica (PSO) e nuvem de partículas encadeadas (CPSO). A BG tradicional difere-se da 1-*st* BG por fazer uma nova busca de parâmetros a cada apresentação de dados, enquanto que a última sempre emprega o par (C,γ) encontrado no primeiro treinamento. No que diz respeito as técnicas de nuvem de partículas, a tradicional PSO faz uma procura de parâmetros a cada novo conjunto, enquanto que a CPSO otimiza a solução de forma encadeada, como uma produção em série, sem reinicializar a nuvem a cada apresentação de dados.

O desempenho do modelo de Kapp, Sabourin e Maupin (2012) foi muito similar ao do PSO tradicional, pois a diferença de erros na fase de teste foi de apenas 0,25%. Já em relação às demais técnicas, a sua performance foi bem superior, pois apresentou respectivamente 25,55%, 48,91% e 22,38% menos erros do que a BG, 1-*st* BG e CPSO. Todavia, dentre todas as metodologias avaliadas, a que encontrou menor número de vetores suporte foi o PSO tradicional. No artigo, não foram realizadas comparações de tempo entre o método proposto e as buscas por *grid* (BG e 1-*st* BG). No entanto, quando comparado às técnicas PSO, observou-se uma diminuição significativa do tempo computacional. Para cinco dos trezes conjuntos avaliados, a redução de Kapp, Sabourin e Maupin (2012) foi superior a uma hora.

Lebrun *et al.* (2008) elaboram um método para selecionar simultaneamente: os dados mais relevantes do conjunto de treinamento, as características desses dados e os parâmetros (C, σ) do SVC, utilizando a busca tabu (BT). Para fazer a simplificação do conjunto de entrada, os autores utilizaram o algoritmo LBG (Linde, Buzo, Gray), que é muito aplicado na área de quantificação vectorial. Desta forma, objetivo da BT era procurar uma solução que englobasse, além das caraterísticas dos dados de treinamento e o par de parâmetros (C, σ), o nível de simplificação k do algoritmo LBG. Para analisar o desempenho do método proposto, os autores avaliaram bases de dados do repositório UCI e um conjunto, denominado *ClassPixel*, constituído por imagens microscópicas de tumores bronquiais. A partir dos resultados evidenciou-se que o modelo foi capaz de reduzir significativamente o número de dados e de características necessárias para o treinamento, proporcionando soluções com baixa quantidade de vetores suporte e taxas de acerto superiores a 81,5%. Neste trabalho, não houveram comparações numéricas com outros métodos de seleção de parâmetros.

Bai, Yang e Zhang (2013) desenvolveram um modelo para selecionar os parâmetros (C, γ) do SVC baseado na técnica *Parallel Artificial Fish Swarm Algorithm* (PAFSA), cuja inspiração se dá no comportamento social de cardumes de peixes na busca de alimentos. O PAFSA difere do tradicional *Artificial Fish Swarm Algorithm* (AFSA) por uma mudança em um de seus *loops*, que evita que o algoritmo se atenha a mínimos locais. Os parâmetros (C, γ) encontrados pelo PAFSA foram empregados na classificação de um problema de reconhecimento de fala e os seus resultados foram comparados com os do tradicional AFSA. A partir dos resultados, evidenciou-se que os parâmetros determinados pelo método proposto proporcionaram ao SVC um aumento médio 3,87% na acurácia. Assim, foi possível reconhecer adequadamente diferentes tipos de palavras, pronunciadas em ambientes com e sem ruído.

LIN *et al.* (2015) criaram um método para determinar as principais características dos dados de treinamento e os parâmetros (C, γ) do SVC, utilizando uma versão modificada do algoritmo *Cat Swarm Optimization*, denominada MCSO-SVM. A técnica *Cat Swarm Optimization* (CSO) baseia-se no comportamento social dos gatos, que simula a sua habilidade de permanecer alerta até mesmo enquanto descansam e a sua forma de rastrear e pegar seus alvos. A modificação que deu origem ao MCSO consiste em uma operação de mutação dos gatos, que propicia ao algoritmo uma melhor busca de soluções. O desempenho do método proposto foi comparado com o da tradicional CSO, utilizando-se conjuntos de dados do repositório UCI. A partir das avaliações, constatou-se que o MCSO-SVM obteve o mesmo desempenho do CSO-SVM em dois dos noves conjuntos analisados. Já nos demais, o MCSO-SVM apresentou uma acurácia, em média, 3,06% maior.

Além da busca linear de parâmetros e dos modelos de seleção que abordam as metaheurísticas (SA, GA, PSO, BT, PAFSA e MCSO), evidenciam-se na literatura técnicas que envolvem expressões analíticas em seu desenvolvimento (VAREWYCK; MARTENS; 2011; WANG; HUANG; CHENG, 2014), meta-aprendizado (GOMES *et al,* 2012; MIRANDA *et al*, 2014) e otimização multi-objetivo (MIRANDA *et al*, 2014). A seguir explicam-se exemplos desses métodos.

Varewyck e Martens (2011) determinaram uma equação, abrangendo um parâmetro de aproximação ψ, que permite encontrar valores aceitáveis para o parâmetro γ do kernel gaussiano, representada por

$$\log_2 \gamma = \log_2 \psi - \log_2 \sigma_c^2 - \log_2 S - 1, \quad (47)$$

onde σ_c^2 é a variância dos vetores pertencentes à classe minoritária e S é a quantidade de características dos dados de treinamento. A partir de alguns testes empíricos e de um procedimento não usual de normalização de dados (passo 1 do QUADRO 4), os autores concluíram que para quase todos os problemas de classificação γ depende apenas da dimensionalidade dos vetores de entrada, cujo valor pode ser heuristicamente calculado por

$$\gamma = \frac{1}{S\sqrt{2}}. \quad (48)$$

Desta forma, o método propõe definir primeiramente o parâmetro γ e depois a constante de regularização C, que por sua vez é selecionada dentro de uma pequena faixa de valores, conforme mostra o procedimento do QUADRO 4.

QUADRO 4 - PSEUDOCÓDIGO DO MÉTODO DE VAREWYCK e MARTENS (2011)

MÉTODO: VAREWYCK e MARTENS (2011)
1. Normalizar os dados de treinamento seguindo os passos 1.1 a 1.5
   1.1. Fazer o histograma de cada atributo (característica) do vetor
   1.2. Encontrar a classe do histograma com mais elementos
   1.3. Calcular a média dos valores pertencentes à classe do passo 1.2
   1.4. Subtrair dos valores originais a média encontrada no passo 1.3
   1.5. Multiplicar o valor obtido no passo 1.4 por um fator que resulte em uma variância igual a 1.
2. Determinar o parâmetro γ pela equação (48)
3. Escolher $\log_2 C = 1$ como constante de regularização e treinar o SVC
4. Executar um segundo treinamento no SVC usando $\log_2 C = 3$. Se a taxa de erro encontrada for superior à verificada no passo 3, realizar um terceiro treinamento com $\log_2 C = -1$.
5. Selecionar como constante de regularização C, dentre os valores dos passos 3 e 4, aquele que resultou na menor taxa de erro.

FONTE: VAREWYCK e MARTENS (2011).

Varewyck e Martens (2011), utilizando o LIBSVM, avaliaram treze conjuntos de dados, sendo oito do repositório UCI, e compararam seus resultados com o da busca por *grid.* Apesar da BG apresentar menores taxas de erros de classificação, o teste estatístico de *Wilcoxon* não evidenciou diferenças significativas entre ambos os métodos. A exceção se deu em apenas um conjunto, no qual a BG obteve 0,92% de

erros contra 10,66% do proposto. Os autores não souberam explicar o motivo pelo qual isso ocorreu.

Wang, Huang e Cheng (2014), inspirados no trabalho de Varewyck e Martens (2011), empregaram a expressão (47) para elaborar seu método de seleção de parâmetros. Porém, ao invés de normalizar os dados de treinamento conforme o passo 1 do QUADRO 4, eles os padronizaram para média igual a zero e variância igual a 1. A partir desse procedimento e da simplificação $\log_2 \psi = 0$, também adotada por Varewyck e Martens (2011), Wang, Huang e Cheng (2014) concluíram que é possível obter o parâmetro γ por meio da relação

$$\gamma = \frac{1}{2S}. \tag{49}$$

Assim, o modelo proposto por Wang, Huang e Cheng (2014), igualmente ao de Varewyck e Martens (2011), define γ antes da constante de regularização C. Para encontrar o valor de C, os autores partiram do princípio que, se for estabelecido um valor apropriado para o parâmetro do kernel e a amostra de treinamento não contiver erros (*outliers*), a solução ótima do SVC $\boldsymbol{\alpha}^*$ será limitada superiormente por um valor finito. Logo, se γ for determinado por (49) e os possíveis erros do conjunto de entrada forem retirados, o valor limite de $\boldsymbol{\alpha}^*$ pode ser visto como a constante de regularização. O QUADRO 5 resume essa ideia, evidenciando os passos do algoritmo de Wang, Huang e Cheng (2014).

QUADRO 5 - PSEUDOCÓDIGO DO MÉTODO DE WANG, HUANG e CHENG (2014)

MÉTODO: WANG, HUANG e CHENG (2014)

1. Padronizar os dados de treinamento para média igual a 0 e variância igual a 1.
2. Definir o parâmetro do kernel gaussiano por (49)
3. Estabelecer a constante de regularização C = 1 e treinar o SVC
4. Obter os vetores suporte provenientes do passo 3
5. Identificar quais vetores suporte são do tipo *Bound* (erro) e retirá-los da amostra de dados.
6. Atribuir um valor relativamente grande para a constante de regularização C (por exemplo C=1024) e treinar o SVC sem os dados eliminados no passo 5.
7. Utilizar o maior valor dos multiplicadores de Lagrange obtidos no passo 6 como constante de regularização C.

FONTE: WANG, HUANG e CHENG (2014).

Para avaliar a eficiência do método proposto, Wang, Huang e Cheng (2014) analisaram oito conjuntos de referência, retirados do repositório UCI e do projeto Statlog. Os quatro menores conjuntos (com menos 2310 dados) foram comparados com a busca por *grid* e os quatro maiores com o método de Varewyck e Martens (2011). Em relação às pequenas amostras, apesar da metodologia proposta ter sido ligeiramente melhor que BG em apenas um caso, as diferenças entre as acurácias não foram superiores a 1,30% em nenhuma das quatro avaliações. Contudo, o tempo despendido pela técnica de Wang, Huang e Cheng (2014) foi em torno de cinquenta vezes menor que o da BG. Todavia, destaca-se que essa comparação aparenta não ser justa uma vez que se utilizou um *grid* de tamanho 9 x 13 contra três avaliações do parâmetro C, feitas por Wang, Huang e Cheng (2014), conforme mostram os passos 3, 6 e 7 do QUADRO 5. Para os grandes conjuntos de dados, o desempenho da proposta foi superior ao de Varewyck e Martens (2011) em metade dos casos (em média 5,00% melhor) e similar nos demais. Já a sua execução foi em média 6,70% mais rápida que a de Varewyck e Martens (2011) em todas as situações.

Um método de seleção de parâmetros do SVC que se diferencia dos anteriormente explicados é o de Miranda *et al*. (2014). Inspirados no trabalho de Gomes *et al.* (2012), que aborda o meta-aprendizado para encontrar parâmetros para o algoritmo de regressão - *Support Vector Regression* (SVR) - Miranda *et al*. (2014) propuseram uma arquitetura híbrida que combina o meta-aprendizado com uma nuvem de partículas multi-objetivo para determinar a constante de regularização C e o parâmetro γ do kernel gaussiano do SVC. No meta-aprendizado (MA), a seleção de parâmetros é tratada como um aprendizado supervisionado, em que cada padrão de treinamento (meta-exemplo) armazena características de problemas antigos e o desempenho de pares (C, γ) candidatos, cujas informações são usadas na previsão de parâmetros do problema atual. Para criar a meta-base (conjunto de meta-exemplos), Miranda *et al.* (2014) utilizaram informações de 100 conjuntos de dados do repositório UCI. Na técnica proposta, as configurações de parâmetros indicadas pela MA foram empregadas como solução inicial para seis algoritmos de nuvem de partículas multi-objetivo: MOPSO, MOPSO-CDR, MOPSO-CDRS, CSS-MOPSO, m-DNPSO e MOPSO-CDLS[6], que tinham por finalidade fornecer parâmetros que

[6] No trabalho de Miranda *et al.* (2014) foram apenas indicadas as siglas dos algoritmos de nuvem de partículas multi-objetivo, sem seus nomes por extenso.

maximizassem a taxa de acerto e minimizassem o número de vetores de suporte. A partir dos resultados, Miranda *et al.* (2014) verificaram que o meta-aprendizado sugeriu boas soluções iniciais para os seis algoritmos de otimização, permitindo que as buscas de parâmetros começassem por regiões mais promissoras do que as dos algoritmos tradicionais (sem o uso do MA), o que favoreceu a convergência do problema e a obtenção de melhores soluções.

A revisão bibliográfica realizada neste trabalho, embora não possa englobar todas as possibilidades existentes, teve por objetivo evidenciar os principais e mais recentes métodos de seleção de parâmetros (C, γ) disponíveis na literatura. Como pôde ser visto, nenhum dos modelos apresentados abordam a técnica *quadtree* em seu desenvolvimento.

Ainda, apesar de nesta tese estudar-se apenas metodologias destinadas ao algoritmo de classificação, SVC, é importante ressaltar que na área de regressão, SVR, também não constam métodos de determinação de parâmetros que envolvam *quadtree.* Nesse campo, somente constatou-se modelos baseados em metaheurísticas, (WANG *et al*, 2005; YUAN, 2012, CHE, 2013), fórmulas empíricas (CHERKASSKY; MA, 2004), limitação do espaço de busca (ORTIZ-GARCIA *et al,* 2009), busca linear (PANG *et al*, 2011) e meta-aprendizado (GOMES *et al,* 2012), por exemplo.

Para finalizar, mesmo sendo menos comuns e frequentes, os métodos que consideram outros kernels na busca de parâmetros do SVC, em especial o polinomial, kernel normalizado e combinação de kernels, também não tratam de *quadtree.* Dentre as poucas abordagens encontradas, destacam-se as que trabalham com algoritmos genéticos (LESMANN; STAHLBOCK; CRONE, 2006), algoritmo cultural caótico adaptativo (GUO; YANG; XIAO, 2010), equações matemáticas e propriedades envolvendo a normalização de kernels (LI *et al,* 2012; LIU; XIU, 2013).

3 *QUADTREE*

A *quadtree* é uma estrutura hierárquica de dados criada pela partição do espaço em um conjunto de quadrantes, cujos lados são representados por potências de dois. Tal técnica foi desenvolvida por Finkel e Bentley (1974), porém muitas das contribuições conferidas ao assunto foram dadas por Hanan Samet, pesquisador e professor da Universidade de Maryland. Dentre suas inúmeras publicações, destacam-se Samet (1981, 1982, 1984, 1990, 1994), referências amplamente citadas nesta tese.

A finalidade deste capítulo é abordar a teoria da *quadtree* e apresentar os algoritmos essenciais à compreensão e execução do método proposto. Primeiramente, introduzem-se as definições básicas de árvores enraizadas e as específicas à *quadtree.* Na sequência, explicam-se o funcionamento da técnica, os critérios de balanceamento e a determinação de vizinhos. Por fim, mostram-se recentes aplicações da *quadtree*, presentes em diferentes áreas de atuação.

3.1 DEFINIÇÕES BÁSICAS

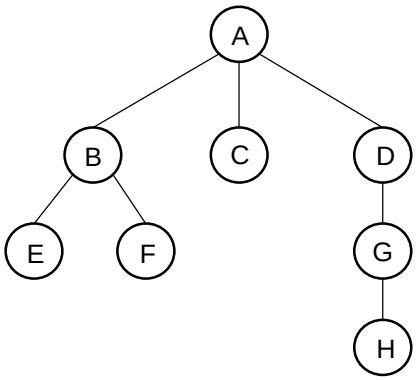
Em ciência da computação, devido a sua vasta empregabilidade, recebe destaque a estrutura de dados denominada árvore. Nesta seção, apresentam-se os conceitos fundamentais relacionados a um de seus tipos, a árvore enraizada.

Uma árvore enraizada é uma árvore livre[7], em que um dos seus vértices diferencia-se dos demais. Esse vértice distinto chama-se raiz da árvore e, a partir dele, existe um único caminho que o conecta até qualquer outro vértice da árvore, também intitulados nós. A existência desse trajeto exclusivo estabelece uma estrutura hierárquica na árvore (VALIENTE, 2010).

Ao considerar um nó x em uma árvore enraizada T de raiz r, qualquer nó y no caminho simples de r a x é chamado ancestral de x. Por consequência, se y é um ancestral de x, então x é um descendente de y (CORMEN *et al*.,2012). A FIGURA 14 ilustra uma árvore enraizada, na qual a sua raiz é denotada pelo nó A.

[7] Uma árvore livre é um grafo acíclico conexo não dirigido. Diz-se que um grafo não dirigido é conexo quando todo vértice pode ser alcançado a partir de todos os outros vértices. Maiores explicações sobre a teoria dos grafos podem ser encontradas em Cormen *et al.* (2012) e Valiente (2010).

FIGURA 14 - REPRESENTAÇÃO DE UMA ÁRVORE ENRAIZADA



FONTE: A autora (2016).

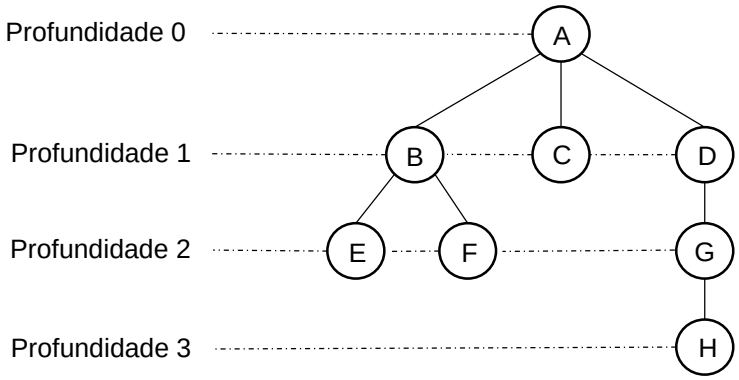
Ainda pela FIGURA 14, tem-se a título de exemplo que os nós D e G são ancestrais do nó H e, consequentemente, H é descendente de D e G. Contudo, a ligação entre D e G é direta enquanto a de D e H é indireta.

Destaca-se que nós diretamente conectados apresentam uma relação especial entre si. De acordo com Cormen *et al*. (2012), em uma árvore T, se a última aresta no caminho simples da raiz r a um nó x é (y, x), então y é pai de x e x é filho de y. Logo, se dois nós possuem o mesmo pai, ambos são considerados irmãos. A raiz é o único nó da árvore que não tem pai.

Distinguem-se ainda os nós que não têm filhos daqueles que os possuem. Os primeiros são chamados folhas ou nós externos, enquanto os segundos são ditos nós internos. Novamente pela FIGURA 14, verifica-se que B é pai de E e F e, portanto, é um nó interno. Já os nós E e F, além de irmãos, são folhas.

Em uma árvore enraizada T, o número de filhos de um nó x é denominado grau de x e o comprimento do caminho simples da raiz r ao nó x é dita profundidade de x em T. Assim, todos os nós que estão na mesma profundidade formam um nível de uma árvore (CORMEN *et al*.,2012). A FIGURA 15 ilustra esses conceitos, mostrando por exemplo que os nós B, C e D têm profundidade 1 e, dessa maneira, estão no mesmo nível.

FIGURA 15 - EXEMPLO DE UMA ÁRVORE COM ÊNFASE NA PROFUNDIDADE DE SEUS NÓS



FONTE: A autora (2016).

Por fim, introduz-se a definição de árvore ordenada, que consiste em uma árvore enraizada na qual os filhos de cada nó estão ordenados. Isto significa que se um nó tem k filhos, então existe um primeiro filho, um segundo filho e, outros na sequência, até o k-ésimo filho (CORMEN *et al*., 2012).

Este trabalho baseia-se num tipo especial de árvore ordenada, intitulada *quadtree.* Nesta árvore, cada nó pai possui quatro filhos ordenados conforme explica-se na próxima seção.

3.2 CONCEITOS GERAIS

Segundo Samet (1984, 1990, 1994), o termo *quadtree* é utilizado para descrever uma classe de estruturas hierárquicas de dados cuja propriedade comum é o princípio da decomposição recursiva do espaço. As *quadtrees* diferenciam-se pelos seguintes critérios:

a) Tipo de dados que representam;

b) Princípio norteador do processo de decomposição; e

c) Resolução (variável ou não).

No que diz respeito aos dados, as *quadtrees* podem ser usadas para representar pontos, retângulos, regiões, curvas, superfícies e volumes. A decomposição pode ser realizada em partes iguais em cada nível, denominada decomposição regular, ou governada pelos dados de entrada. Já a resolução, que consiste no número de vezes que o processo de decomposição é executado, pode ser fixada *a priori* ou depender das propriedades dos dados de entrada.

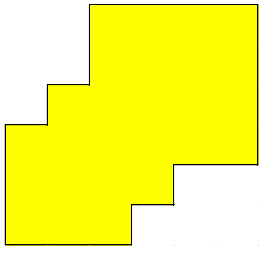
Dentre as abordagens da *quadtree*, a mais comumente utilizada é a de representação de regiões (*region quadtree*), que é a estudada na presente pesquisa. Desta forma, os conceitos explicados a partir desta seção referem-se unicamente a essa aplicação.

3.3 FUNCIONAMENTO DA *QUADTREE*

O princípio da técnica *quadtree* baseia-se na divisão sucessiva do espaço em quatro quadrantes de mesmo tamanho. A ideia fundamental é, ao realizar a fragmentação, identificar pelas suas intersecções quais quadrantes estão inteiramente contidos na área de interesse, parcialmente inseridos ou vazios. Aqueles que estiverem parcialmente contidos, ou seja que possuírem dados internos e externos à região considerada, são recursivamente divididos em novos quadrantes até que todos esses se tornem homogêneos (possuidores de apenas dados externos ou internos). Quando essa condição é atingida, encerra-se o processo de divisão.

A FIGURA 16 mostra, à esquerda, um exemplo de região investigada pela *quadtree* e, à direita, a representação binária do espaço denotada por uma matriz 2^3 x 2^3. Observa-se pela FIGURA 16 (à direita) que os números “1” indicam os elementos de interesse da figura, interiores a região analisada, enquanto que os “0” correspondem aos elementos externos.

FIGURA 16 - REGIÃO E SUA CORRESPONDENTE MATRIZ BINÁRIA



0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1
0	0	0	0	1	1	1	1
0	0	0	1	1	1	1	1
0	0	1	1	1	1	1	1
0	0	1	1	1	1	0	0
0	0	1	1	1	0	0	0

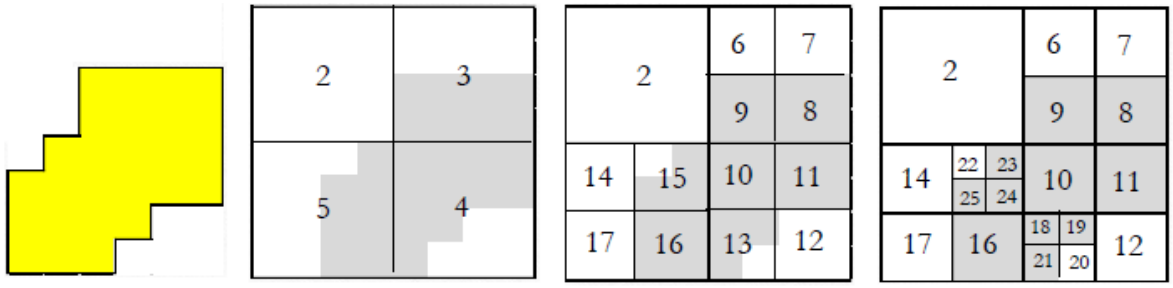
FONTE: Adaptado de SAMET (1984, 1990, 1994).

Para o exemplo da FIGURA 16, à medida que a partição do espaço é realizada, se os quadrantes gerados possuírem tanto elementos “1” quanto “0”, eles devem ser novamente divididos até que todos os seus subquadrantes atinjam a condição de homogeneidade. Desta forma, considera-se a resolução da *quadtree* de

regiões como variável, pois ela depende dos dados de entrada (SAMET, 1984, 1990, 1994).

A FIGURA 17 ilustra, em detalhes, todo o processo de funcionamento da *quadtree* para o exemplo da FIGURA 16. De acordo com Greaves e Borthwick (1999), existem inúmeros sistemas de numeração usados para armazenar as informações da *quadtree.* Em Samet (1981, 1982) e Finkel e Bentley (1974) encontram-se dois deles. Contudo, na FIGURA 17 apresenta-se a execução das divisões conforme o sistema estabelecido na presente tese, que foi elaborado com o objetivo de facilitar a implementação computacional do método proposto. Logo, neste trabalho, os quadrantes são numerados segundo a sua localização (direção), atendendo a seguinte ordem: noroeste (NO), nordeste (NE), sudeste (SE) e sudoeste (SO). A FIGURA 18 aponta essa sequência estipulada.

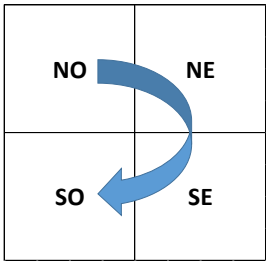
FIGURA 17 - FUNCIONAMENTO DA *QUADTREE*



(a) (b) (c) (d)

FONTE: A autora (2016).

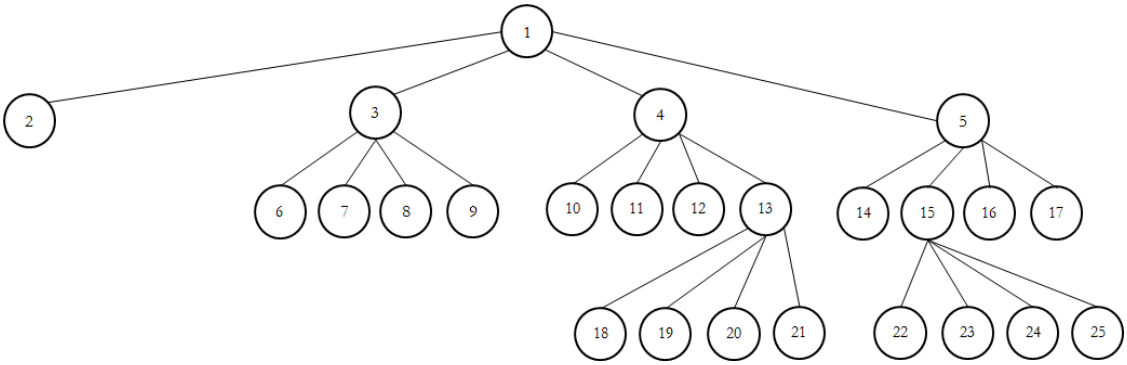
FIGURA 18 - SISTEMA DE NUMERAÇÃO ADOTADO



FONTE: A autora (2016).

Ressalta-se que todo o processo demonstrado na FIGURA 17 é representado por uma árvore ordenada de grau 4, conforme mostra a FIGURA 19. Nessa representação, a raiz contém todas as informações referentes às divisões realizadas e cada quadrante gerado caracteriza um nó da árvore.

FIGURA 19 - ÁRVORE *QUADTREE*



FONTE: A autora (2016).

Na terminologia da *quadtree*, os quadrantes que contêm somente dados internos à região são denominados PRETOS e os que possuem apenas dados externos são BRANCOS. Os nós correspondentes a esses dois tipos de quadrantes tornam-se folhas, significando que nenhuma divisão a mais será necessária. Já os quadrantes heterogêneos, denominados CINZA, são nós internos e subdivididos até que todos os seus filhos virem folhas (SAMET, 1984, 1990, 1994).

Apresentados esses conceitos, explica-se a seguir o processo de divisão e de numeração das FIGURAS 17 e 18. Analisando a FIGURA 17b, percebe-se que após a fragmentação inicial da *quadtree* todos os quadrantes gerados, com exceção do número 2, são heterogêneos. Portanto, deve-se continuar a execução da técnica fracionando-os. Obedecendo ao sistema de numeração convencionado, particiona-se primeiramente o quadrante 3, seguido do 4 e por fim o 5. Assim, o ordenamento dos futuros quadrantes se inicia a partir dos filhos de 3, que recebem os números 6 (NO), 7 (NE), 8 (SE) e 9 (SO). Na sequência, numeram-se respectivamente os filhos de 4 e 5, chegando-se a quantidade de 17.

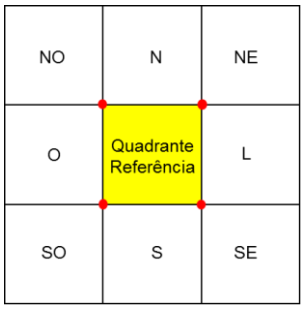
Finalizada essas divisões, observa-se pela FIGURA 17c que apenas os quadrantes de número 13 e 15 são heterogêneos. Todos os demais tornaram-se folhas. De acordo com a nomenclatura da *quadtree*, tem-se que os nós 6, 7, 12, 14 e 17 são BRANCOS enquanto que os nós 8, 9, 10, 11 e 16 são PRETOS. Logo, somente os nós CINZA, 13 e 15, devem ser divididos. Pela ordem definida, particiona-se primeiramente o 13, gerando os subquadrantes 18, 19, 20 e 21, e depois o 15, criando-se os nós 22 ao 25.

Terminada esta etapa, verifica-se pela FIGURA 17d que todos os quadrantes resultantes são homogêneos. Assim, encerra-se o funcionamento da *quadtree* do qual sucede uma árvore de profundidade 3, conforme ilustrado na FIGURA 19.

3.4 *QUADTREE* BALANCEADA

Em uma *quadtree*, um quadrante ou nó pode possuir no máximo oitos vizinhos de mesmo tamanho: quatro de aresta, nas direções norte (N), leste (L), sul (S) e oeste (O), e quatro de vértice, nas direções noroeste (NO), nordeste (NE), sudeste (SE) e sudoeste (SO). A FIGURA 20 apresenta essas possibilidades.

FIGURA 20 - POSSÍVEIS VIZINHOS DE UM QUADRANTE

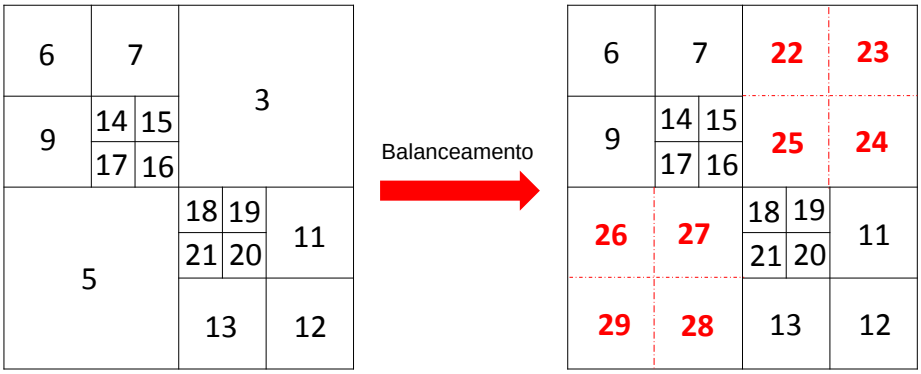


FONTE: Adaptado de FRANCISQUETTI (2010).

Diz-se que oito é a quantidade máxima, uma vez que o quadrante pode estar localizado nos limites do espaço considerado. Por exemplo, retomando a FIGURA 17b, verifica-se que o quadrante 3 tem apenas três vizinhos de mesmo tamanho: quadrante 2 (a oeste), o 4 (ao sul) e o 5 (a sudoeste). Nas demais direções, ele não os possui, pois está na fronteira do espaço de busca.

Para melhorar a convergência de um problema envolvendo *quadtree*, é desejável que a mesma esteja balanceada. Isso significa que a maior diferença entre os níveis de nós adjacentes não pode exceder a 1 para os vizinhos de aresta e a 2 para os de vértice. Segundo Moore (1995), a condição de balanceamento previne mudanças abruptas nos tamanhos dos elementos da malha. Consequentemente, isso limita a variação na disposição dos quadrantes, facilitando a solução (GREAVES; BORTHWICK, 1999). A FIGURA 21 destaca a diferença entre uma *quadtree* balanceada (à direita) e outra não (à esquerda).

FIGURA 21 - *QUADTREE* NÃO BALANCEADA VERSUS BALANCEADA



FONTE: A autora (2016).

O balanceamento de uma *quadtree* é realizado somente após o término de todo o seu processo de divisão, ou seja, quando não há mais iterações a serem feitas. Para isso, deve-se avaliar apenas os nós folhas da árvore, observando quais são seus vizinhos e as diferenças de níveis entre eles. Se a diferença exceder aos critérios estabelecidos, deve-se dividir o nó analisado. O QUADRO 6 expõe o pseudocódigo do algoritmo de balanceamento.

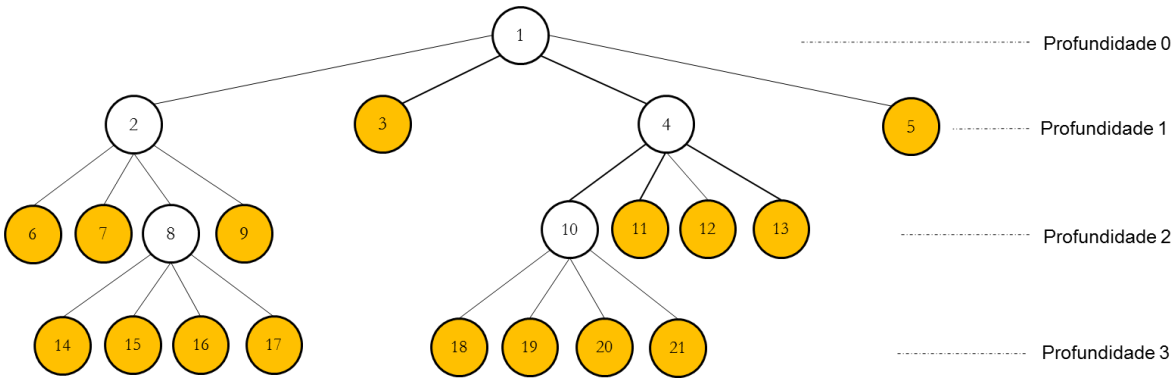
QUADRO 6 - PSEUDOCÓDIGO DO ALGORITMO DE BALANCEAMENTO DA *QUADTREE*

```
PROCEDIMENTO: BALANCEAMENTO
1. Inserir todos os nós folhas NO_i em uma lista L
2. ENQUANTO L não for vazia
3.   PARA cada NO_i
4.     PARA cada direção de aresta ∈ { N, E, S, O}
5.       SE existir nó vizinho na direção avaliada, de mesmo tamanho ou maior, ENTÃO
6.         SE o nó vizinho tiver filhos ENTÃO
7.           SE os filhos do nó vizinho, cujos lados são adjacentes ao do NO_i, tiverem filhos ENTÃO
8.             Dividir NO_i
9.             Inserir os quatro filhos de NO_i em L
10.            Retornar ao passo 2
11.  SE nenhum balanceamento por aresta foi realizado ENTÃO
12.    PARA cada direção de vértice ∈ { NO, NE, SE, SO}
13.      SE existir nó vizinho na direção avaliada, de mesmo tamanho ou maior, ENTÃO
14.        SE o nó vizinho tiver netos ENTÃO
15.          SE o neto do nó vizinho, cujo vértice é adjacente ao vértice do NO_i, tiver filhos ENTÃO
16.            Dividir NO_i
17.            Inserir os quatro filhos de NO_i em L
18.            Retornar ao passo 2
```

FONTE: Adaptado de Paiva Neto (2015).

Para melhor compreender o funcionamento do algoritmo apresentado no QUADRO 6, ilustra-se a *quadtree* não balanceada da FIGURA 21 (à esquerda) por meio da sua representação de árvore, denotada pela FIGURA 22. Percebe-se pela FIGURA 22 que os nós destacados em laranja são as folhas e, portanto, os que devem ser avaliados ao executar o processo de balanceamento. Desta forma, ao aplicar o algoritmo do QUADRO 6, tem-se que a lista L é dada por {3, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21}. Assim, o primeiro nó a ser analisado é o de número 3.

FIGURA 22 - ÁRVORE *QUADTREE* NÃO BALANCEADA



FONTE: A autora (2016).

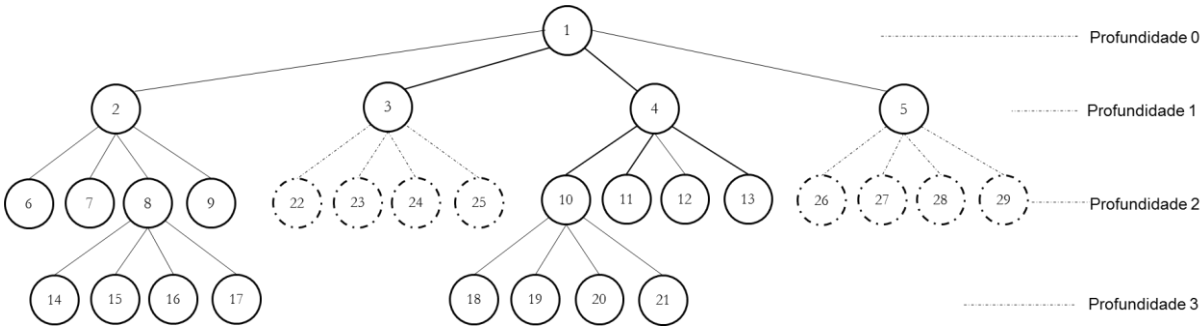
Pela FIGURA 21 (à esquerda), no que diz respeito aos vizinhos de aresta, constata-se que o nó 3 não possui quadrantes adjacentes nas direções norte e leste, o que implica na investigação da aresta sul. Nessa direção, atendendo ao requisito de busca de vizinhos de mesmo tamanho ou maior, onde a procura do primeiro tem prioridade, depara-se com o nó 4 como o vizinho de igual dimensão.

Pelas FIGURAS 21 (à esquerda) e 22, verifica-se que o 4 é um nó interno e, dentre seus filhos, os do tipo NO e NE, numerados respectivamente por 10 e 11, são os que possuem lado adjacente ao nó 3. Logo, pelo passo 7 do QUADRO 6, deve-se avaliar se os nós 10 e 11 têm filhos. Como o quadrante 10 é pai, é preciso balancear a *quadtree* fracionando o nó 3. De acordo com o sistema aqui convencionado, os filhos de 3 são enumerados de 22 a 25 e inseridos na lista L para futura averiguação.

Uma vez efetuado o balanceamento do nó 3, por critério de aresta, dispensa-se a localização dos seus vizinhos de vértice. Assim, dá-se continuidade ao algoritmo, examinando a próxima folha de L, que neste caso é a de número 5.

Procedendo de forma análoga para o nó 5 e demais folhas de L, chega-se como resposta na *quadtree* balanceada da FIGURA 21 (à direita), representada pela árvore da FIGURA 23.

FIGURA 23 - ÁRVORE *QUADTREE* BALANCEADA



FONTE: A autora (2016).

Na FIGURA 23, os nós destacados pelas linhas tracejadas, correspondentes aos filhos dos quadrantes 3 e 5, são os gerados após o balanceamento. Nota-se, ao comparar as árvores das FIGURAS 22 e 23, que na *quadtree* não balanceada existem folhas vizinhas de aresta com diferença de nível igual a dois, o que ultrapassa o limite desejado. São exemplos disso: o nó 3 e seus vizinhos 18 e 19; e o nó 3 e os quadrantes 15 e 16. Já na versão balanceada, isso não mais ocorre, pois nenhuma diferença entre as folhas é superior a um.

3.5 DETERMINAÇÃO DE VIZINHOS

A tarefa de determinar os vizinhos de um quadrante é ponto crucial no trabalho com *quadtrees*, principalmente no que diz respeito ao balanceamento da árvore. Conforme visto, só é possível balanceá-la uma vez que se conheça os nós adjacentes das folhas.

Em Samet (1981, 1982) consta um método muito eficiente para localizar vizinhos em uma *quadtree*. O grande diferencial dessa técnica é que ela não depende do sistema de numeração estipulado, nem das coordenadas dos nós e muito menos do tamanho dos quadrantes (FRANCISQUETTI, 2010). Portanto, por ser um método mais geral, optou-se por adotá-lo na presente tese.

O método de Samet (1981, 1982) baseia-se na localização do ancestral comum mais próximo. O seu entendimento depende de algumas funções e operações,

que envolvem blocos de quadrantes e suas fronteiras (arestas e vértices), bem como notações importantes. O QUADRO 7 apresenta as funções de interesse, onde P é um nó e I é um tipo de quadrante.

QUADRO 7 - FUNÇÕES NECESSÁRIAS AO ALGORITMO DO VIZINHO

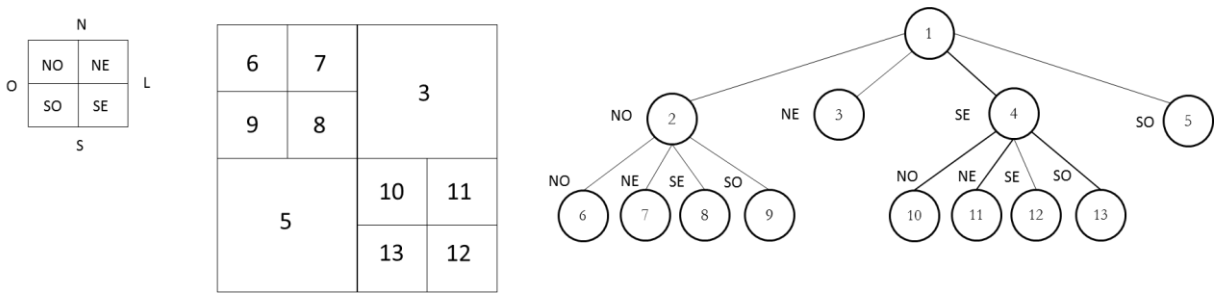
FUNÇÃO	DESCRIÇÃO
PAI (P)	Retorna o pai do nó P ou nulo (caso P seja a raiz da árvore).
FILHO (P, I)	Retorna o filho do nó P localizado no quadrante tipo I.
TIPO_FILHO (P)	Retorna o tipo de filho (rótulo) ao qual o nó P está associado, que pode ser: NO, NE, SE ou SO.
TIPO_NÓ (P)	Retorna o tipo do nó conforme seu conteúdo: BRANCO, CINZA ou PRETO. Em outras palavras, verifica se os nós são folhas (BRANCO ou PRETO) ou internos (CINZA).

FONTE: SAMET (1981, 1982).

Para esclarecer o uso do QUADRO 7, aplicam-se as funções nele descritas no exemplo da FIGURA 24. A FIGURA 24 ilustra, da esquerda para a direita: i) a orientação dos quadrantes e a direção das suas arestas, ii) um exemplo de malha *quadtree* e; iii) sua respectiva árvore. Desta forma, para essa situação, observa-se que:

- PAI (6) = 2; ou seja, a função identifica que o nó 2 é o pai de 6.
- FILHO (4, NE) = 11; reconhece que o filho de 4, localizado no seu quadrante NE, é o nó 11.
- TIPO_FILHO (5) = SO; verifica que o nó 5 é um filho do tipo SO, pois está situado no quadrante SO do bloco ao qual pertence.
- TIPO_NÓ (2) = CINZA; constata que o nó 2 é interno e, portanto, CINZA.

FIGURA 24 - EXEMPLO DE *QUADTREE* PARA ELUCIDAR O USO DAS FUNÇÕES E OPERAÇÕES



FONTE: A autora (2016).

Samet (1981, 1982) define, além das funções anteriormente explicadas, outras quatro operações essenciais à mobilidade entre os blocos, expostas no QUADRO 8. Destaca-se que cada uma dessas operações está vinculada a uma matriz particular, conforme será apresentado. No QUADRO 8, a notação B denota uma fronteira, também entendida por direção.

QUADRO 8 - OPERAÇÕES UTILIZADAS NO ALGORITMO DO VIZINHO

OPERAÇÃO	**DESCRIÇÃO**
ADJACENTE (B, I)	Retorna verdadeiro (V) se, e somente se, o quadrante tipo I é adjacente à fronteira B do bloco ao qual pertence. Caso contrário, retorna falso (F).
REFLETE (B, I)	Retorna o rótulo do bloco de igual tamanho, que é adjacente à direção B do quadrante tipo I.
LADO_COMUM (Q1, Q2)	Retorna a aresta do bloco em que os quadrantes Q1 e Q2 estão contidos, que é comum a eles. Se Q1 e Q2 não forem irmãos adjacentes ou tratarem-se do mesmo quadrante, então o valor retornado é indefinido.
QUADRANTE_OPOSTO (Q)	Retorna o tipo do quadrante, dentro do mesmo bloco, que não tem aresta comum com o quadrante Q.

FONTE: SAMET (1981, 1982).

A compreensão das operações do QUADRO 8 torna-se mais fácil quando relacionadas à *quadtree* da FIGURA 24. Assim, iniciando pela relação de adjacência, tem-se por exemplo que:

- ADJACENTE (O, SO) = V. Percebe-se pelo desenho esquemático das direções e orientações dos quadrantes, à esquerda da FIGURA 24, que o quadrante SO é adjacente a aresta oeste (O). Essa condição é atendida por qualquer nó SO da FIGURA 24. Assim, verifica-se que os nós 5, 9 e 13 são desse tipo e, consequentemente, adjacentes a aresta O do bloco ao qual pertencem.

Todas as relações de adjacência, ocorridas entre direções e quadrantes, estão descritas na matriz exibida no QUADRO 9. Nessa matriz, as linhas e colunas correspondem respectivamente a direção (fronteira) B e ao tipo de quadrante I.

QUADRO 9 - MATRIZ DE ADJACÊNCIA

Quadrante / Direção	NO	NE	SE	SO
N	V	V	F	F
L	F	V	V	F
S	F	F	V	V
O	V	F	F	V

FONTE: SAMET (1981, 1982).

A operação REFLETE (B, I) espelha o quadrante I por meio da aresta B. A resposta desse espelhamento é sempre um bloco de mesmo tamanho, que pode ser ou não irmão de I. As condições de reflexão, estabelecidas na matriz do QUADRO 10, são válidas para quaisquer nós da árvore. Tal qual a de adjacência, as linhas e colunas da matriz de reflexão indicam, nessa ordem, a direção B e o tipo de quadrante I. Logo, tem-se como exemplo:

- REFLETE (S, NE) = SE. Novamente pela FIGURA 24, nota-se que os nós 3, 7 e 11 são do tipo NE. Se esses forem refletidos por meio da sua aresta sul (S) obtém-se respectivamente como solução: os nós 4, 8 e 12, os quais são SE. Verifica-se, para as três situações avaliadas, que os quadrantes encontrados possuem a mesma dimensão do original e, nesses casos, são seus irmãos.

QUADRO 10 - MATRIZ DE REFLEXÃO

QUADRANTE / DIREÇÃO	NO	NE	SE	SO
N	SO	SE	NE	NO
L	NE	NO	SO	SE
S	SO	SE	NE	NO
O	NE	NO	SO	SE

FONTE: SAMET (1981, 1982).

A operação LADO_COMUM (Q1, Q2) se aplica a dois quadrantes Q1 e Q2 pertencentes ao mesmo bloco. Baseando-se na matriz pertinente, representada pelo QUADRO 11, constata-se que:

- LADO_COMUM (NO, NE) = N. Pela FIGURA 24, fica fácil visualizar que os nós 6 e 7, por exemplo, são do tipo NO e NE e têm em comum a aresta N do nó 2, que é o bloco ao qual pertencem. De mesmo modo, isso ocorre

para os nós 2 e 3, que compartilham o lado N da raiz, e para o 10 e 11, que dividem a aresta norte do nó 4.

- LADO_COMUM (SE, SE) = Ω. Neste caso, como Q1 e Q2 referem-se ao mesmo quadrante, a resposta da operação é indefinida. Pela FIGURA 24, isso aconteceria caso o nó 8, de tipo SE, fosse comparado com ele mesmo, tratando-se de uma indefinição.
- LADO_COMUM (SO, NE) = Ω. Nesta situação, como os quadrantes de rótulos SO e NE não são irmãos adjacentes por aresta (tocam-se apenas por vértice), o retorno da operação é indefinido. Pela FIGURA 24, observa-se que os nós 3 (NE) e 5 (SO) não compartilham aresta comum da raiz.

QUADRO 11 - MATRIZ DE ARESTA COMUM

Q1 \ Q2	**NO**	**NE**	**SE**	**SO**
NO	Ω	N	Ω	O
NE	N	Ω	E	Ω
SE	Ω	E	Ω	S
SO	O	Ω	S	Ω

FONTE: SAMET (1981, 1982).

Por fim, a quarta operação é a de oposição a um quadrante, que está vinculada à matriz demonstrada no QUADRO 12. Como exemplo da sua aplicação, tem-se que:

- QUADRANTE_OPOSTO (NO) = SE. Pela FIGURA 24, repara-se, por exemplo, que o nó 6 (NO) não tem aresta comum ao 8 (SE), portanto ambos são tidos como opostos. Relembra-se que essa operação é válida somente para quadrantes contidos no mesmo bloco.

QUADRO 12 - MATRIZ DE OPOSIÇÃO

QUADRANTE	**QUADRANTE_OPOSTO (Q)**
NO	SE
NE	SO
SE	NO
SO	NE

FONTE: SAMET (1981, 1982).

Apresentadas as funções e operações estipuladas por Samet (1981, 1982), é possível explicar os algoritmos de procura de vizinhos por ele desenvolvidos, sendo um destinado aos vizinhos de aresta e outro aos de vértice. Uma vez que o segundo depende do primeiro, inicia-se pela explanação do de aresta.

O algoritmo descrito no QUADRO 13, denominado originalmente por GTEQUAL_ADJ_NEIGHBOR (P, D), é capaz de encontrar vizinhos nas direções horizontal e vertical, com tamanho maior ou igual ao do nó avaliado. Desta forma, ele localiza vizinhos que estão em um nível superior ou igual ao do nó de entrada.

Tomando novamente a FIGURA 24 como exemplo, o primeiro caso é ilustrado pela busca do vizinho leste de 8, que resulta no nó 3 como solução, de tamanho e nível superior. Já o segundo caso é exemplificado pela procura do vizinho sul do quadrante 10, da onde localiza-se o nó 13, de mesma dimensão e nível. Em situações especiais, caracterizadas pela ausência de vizinhos na direção solicitada, devido ao nó de entrada ser fronteira ou tratar-se da raiz da árvore, o algoritmo de Samet (1981, 1982) retorna zero como resposta.

Expostos esses detalhes, apresenta-se o seu pseudocódigo no QUADRO 13, onde: o nó P e a direção D são as entradas; e o nó Q, calculado no decorrer do procedimento, está relacionado à saída do problema. Percebe-se, pelo passo 2 do QUADRO 13, que o algoritmo consiste em uma função recursiva, pois faz referência a si mesmo.

QUADRO 13 - ALGORITMO PARA ENCONTRAR VIZINHOS DE ARESTA

```
PROCEDIMENTO: MAIOR_IGUAL_VIZINHO_ADJACENTE (P, D)
1.  SE PAI (P) ≠ 0 E ADJACENTE (D, TIPO_FILHO (P)) = V ENTÃO
2.      Q = MAIOR_IGUAL_VIZINHO_ADJACENTE (PAI (P), D)
3.  SENÃO
4.      Q = PAI (P)
5.  SE Q ≠ 0 E TIPO_NO (Q) = CINZA ENTÃO
6.      VIZINHO DE ARESTA = FILHO (Q, REFLETE (D, TIPO_FILHO (P)))
7.  SENÃO
8.      VIZINHO DE ARESTA = Q
```

FONTE: SAMET (1981, 1982).

O algoritmo de vizinhos de vértice, conhecido por GTEQUAL_CORNER_NEIGHBOR (P, C), busca vizinhos nas direções diagonais do

nó avaliado. Tal qual ao de aresta, ele encontra nós adjacentes de maior ou igual tamanho e fornece zero como solução quando não há vizinhos no sentido pesquisado.

Pela FIGURA 24, tem-se, por exemplo, que os vizinhos do nó 11 nas direções NO e SO são respectivamente os quadrantes 3 e 13, de maior e igual dimensão. Já em NE e SE, o nó 11 não possui adjacentes, pois está na fronteira da figura.

Conforme mencionado, o algoritmo de vizinho de vértice depende do procedimento de aresta (QUADRO 13) para seu funcionamento. Assim como o outro, ele também é recursivo pois, além de chamar a função de aresta, demanda a si mesmo durante a sua execução. Isso pode ser verificado nos passos 3 e 5 do pseudocódigo descrito no QUADRO 14.

Neste algoritmo, as entradas são dadas pelo nó P e pelo quadrante C, que é entendido como a direção em que se procura o vizinho. Já o nó Q é novamente uma variável calculada durante da função, que está diretamente ligada à sua saída.

QUADRO 14 - ALGORITMO PARA ENCONTRAR VIZINHOS DE VÉRTICE

```
PROCEDIMENTO: MAIOR_IGUAL_VIZINHO_VÉRTICE (P, C)
1. SE PAI (P) ≠ 0 E TIPO_FILHO (P) ≠ QUADRANTE_OPOSTO (C) ENTÃO
2.    SE TIPO_FILHO (P) = C ENTÃO
3.       Q = MAIOR_IGUAL_VIZINHO_VÉRTICE (PAI (P), C)
4.    SENÃO
5.       Q = MAIOR_IGUAL_VIZINHO_ADJACENTE (PAI (P), LADO_COMUM (TIPO_FILHO (P), C))
6. SENÃO
7.     Q = PAI (P)
8. SE Q ≠ 0 E TIPO_NO (Q) = CINZA ENTÃO
9.      VIZINHO DE VÉRTICE = FILHO (Q, QUADRANTE_OPOSTO (TIPO_FILHO (P)))
7.  SENÃO
8.      VIZINHO DE VÉRTICE = Q
```

FONTE: SAMET (1981, 1982).

Assim, fazendo-se uso dos algoritmos dos QUADROS 13 e 14 é possível realizar o balanceamento da *quadtree*, cujo método foi apresentado no QUADRO 6 deste capítulo. O balanceamento, segundo explicado, equilibra o tamanho dos quadrantes da malha, facilitando a solução do problema.

3.6 APLICAÇÕES DA *QUADTREE*

A *quadtree* é normalmente empregada no campo do processamento de imagens, computação gráfica e sistemas de informação geográfica. Nesta seção, destacam-se algumas das pesquisas evidenciadas na área, que tratam dessas e outras aplicações.

Hosaka, Kobayashi e Otsu (2009) estudaram a técnica de composição de imagens "*Image matting",* que é muito usada na edição de fotos, vídeos e produção cinematográfica. A mesma consiste em extrair o objeto do primeiro plano de uma imagem e posteriormente combiná-lo com outro plano de fundo, de forma que a composição pareça natural. Assim, para cada pixel da imagem, necessita-se estimar o grau de transparência do objeto e a sua cor original. Para tal, os autores utilizaram o algoritmo *Support Vector Classification* (SVC), para aprimorar a discriminação dos dois planos, precedido do uso da *quadtree.* Neste caso, a *quadtree* foi empregada para gerar linhas guias, chamada "*strokes*", que diferenciam o primeiro plano do de fundo. A FIGURA 25 ilustra um exemplo de imagem, original (à esquerda) e guiada (à direita), onde as linhas vermelhas referem-se ao primeiro plano e as azuis ao de fundo.

FIGURA 25 - FIGURA ORIGINAL E FIGURA COM AS LINHAS "*STROKES*"

FONTE: HOSAKA, KOBAYASHI E OTSU (2009).

Para este problema, ressalta-se, a título de exemplo, que a seleção de parâmetros do SVC foi realizada de forma empírica. A justificativa dos autores é que a busca por *grid,* por ter alto tempo computacional, torna-se impraticável para o problema de *Image matting.* Apesar dessa escolha, Hosaka, Kobayashi e Otsu (2009) encontraram resultados de alta qualidade e bastante satisfatórios quando comparados com algoritmos referências da área. A FIGURA 26 apresenta alguns desses resultados.

FIGURA 26 - RESULTADOS DO *IMAGE MATTING* DE HOSAKA, KOBAYASHI E OTSU (2009)

FONTE: HOSAKA, KOBAYASHI E OTSU (2009).

Jing, Li e Li (2011) utilizaram a *quadtree* combinada as técnicas: filtro de Garbor, análise de imagem de fundo e decomposição *Wavelet*, para inspecionar defeitos na fabricação de tecidos. A *quadtree*, neste caso, tinha como função determinar a específica localização da falha, a partir de uma imagem do produto, conforme mostra a FIGURA 27.

FIGURA 27 - LOCALIZAÇÃO DE DEFEITO EM TECIDO

FONTE: JING, LI E LI (2011).

Popinet (2011) desenvolveu um modelo adaptativo de *quadtree* integrado a um sistema de banco de dados de batimetria (medições de profundidades de corpos hídricos) para simular o espraiamento[8] da onda e a propagação do tsunami ocorrido no oceano Índico em 2004. Em continuidade a sua pesquisa, Popinet (2012) aplicou o mesmo método para analisar o comportamento do tsunami do Japão de 2011.

Fundamentando-se em Popinet (2011, 2012), Tsai *et al.* (2013) usaram um modelo adaptativo de *quadtree* para avaliar o espectro das ondas formadas durante os furacões Katrina e Rita. Uma vez que os modelos utilizados nessa área normalmente empregam malhas estáticas, os autores adotaram uma versão, usando a *quadtree,* que acompanha o movimento dos furacões. O grande ganho do método

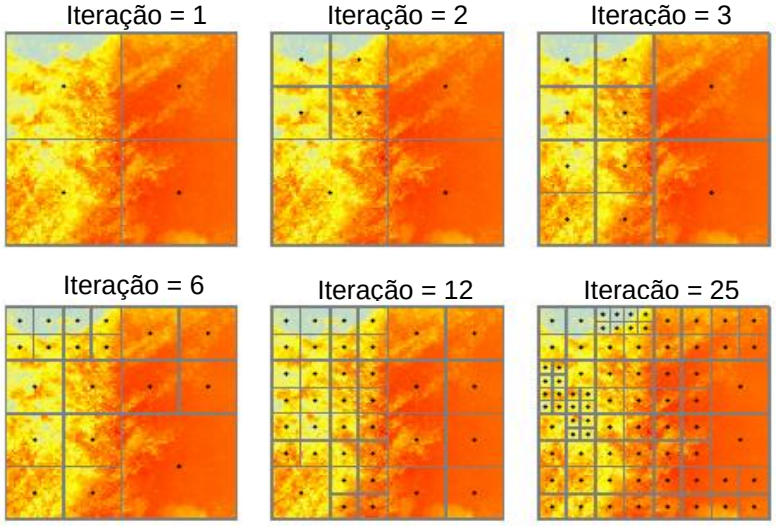
[8] Movimentação ascendente da água da onda incidente, após a rebentação, através da face da praia Fonte: http://www.aprh.pt/rgci/glossario/espraiodaonda.html.

proposto foi a redução no tempo computacional das simulações realizadas. Para o Katrina, enquanto o modelo adaptativo de *quadtree* gastou apenas 3 horas para avaliar os dez dias de passagem do furacão, o de malha estática (*grid* uniforme) necessitou de 13 vezes mais.

Ainda na área de mecânica dos fluídos, Francisqueti (2010) estudou a geração de malhas *quadtrees* aplicadas à simulação de escoamentos. No seu trabalho, a autora analisou o comportamento de diversos tipos de malhas como as não uniformes e as de diferentes refinamentos e interpolações. Já An e Yu (2012) propuseram um modelo adaptativo de *quadtree*, usando a técnica de volume cortado em superfícies curvas ou complexas, para simular o regime de escoamento laminar. O interesse em sistemas adaptativos, na área de dinâmica dos fluídos, se deve ao ajuste contínuo da resolução da malha às características do fluxo.

Yang e Reindl (2015) utilizaram a *quadtree* para determinar o melhor posicionamento de estações de monitoramento de irradiância solar para energia renovável. Para isso, avaliaram uma pequena amostra do banco de dados norte-americano SUNY, correspondente aos valores de irradiância solar incidentes sobre o estado do Colorado e suas proximidades, entre os anos de 2004 e 2005. O algoritmo, por eles proposto, funcionava com base nas variações de irradiância constatadas entre uma região e outra. Desta forma, as divisões da *quadtree* eram executadas até se obter quadrantes (áreas) com um certo grau de similaridade, conforme mostra a FIGURA 28. A partir da delimitação das regiões (clusterização), era possível identificar a quantidade de estações a serem instaladas.

FIGURA 28 - EVOLUÇÃO DA *QUADTREE* NA AVALIAÇÃO DE IRRADIÂNCIA SOLAR



FONTE: YANG E REINDL (2015).

Muhsin *et al.* (2014) explicam que as imagens possuem diversas informações, porém poucas delas são desejáveis a um posterior processamento. Portanto, antes de se trabalhar com uma imagem é importante identificar qual é a sua região de interesse. Neste contexto, os autores aplicaram a *quadtree* como um pré-processamento de imagens, visando segmentar as porções mais heterogêneas de uma figura (cujos *pixels* não possuíam a mesma cor) das demais. Desta forma, por meio da *quadtree*, eles alocaram menos *bits* às regiões de menor interesse (homogêneas) e mais *bits* àquelas de maior valor e informação (heterogêneas), reduzindo o número de partes da imagem a serem transmitidas à etapa de processamento.

Yuen, Lui e Wong (2013) elucidam que o uso da internet e a distribuição de grandes volumes de multimídia requerem formas de comprimir essas informações. Neste contexto, os autores esclarecem que a codificação fractal de imagens proporciona uma maior taxa de compressão que o formato JPEG, porém com menor eficiência de codificação, tornando-a inadequada para muitas aplicações. A fim de superar essas limitações, os autores propuseram uma estrutura progressiva de codificação fractal baseada em *quadtree.* Para tal, utilizou-se uma *quadtree* balanceada, de forma a evitar que a diferença de níveis entre os nós afetasse negativamente a transmissão de *bits* e os efeitos no contraste e brilho. Os resultados obtidos por meio da estrutura proposta mostraram que a qualidade das imagens reconstruídas com o uso da *quadtree* foi superior ao atingido pela tradicional codificação fractal.

4 PROCEDIMENTOS METODOLÓGICOS

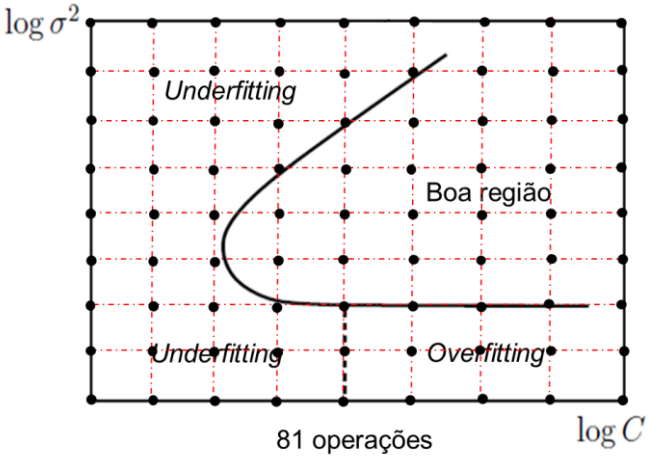
Neste capítulo, descrevem-se os procedimentos metodológicos empregados na presente tese. Para isso, inicia-se pela explicação da ideia norteadora do método *grid-quadtree* que, além de motivar o seu desenvolvimento, embasa todo o funcionamento da proposta. Na sequência, apresentam-se as considerações adotadas referentes a: indicadores de performance, critério de divisão da *quadtree*, balanceamento e criação de uma etapa de refinamento, cuja execução é indispensável para a convergência do problema estudado. Expostos esses fundamentos, detalham-se as características do método propriamente dito, como: a sua forma de inicialização, seus critérios de parada e seu pseudocódigo. Por fim, explicam-se os dados utilizados e a maneira como serão validados os resultados desta pesquisa.

4.1 IDEIA NORTEADORA DO MÉTODO *GRID-QUADTREE*

O método *grid-quadtree* (GQ) de determinação de parâmetros do SVC, proposto neste trabalho, foi implementado por meio da linguagem de programação VB.net e tem por objetivo localizar um bom par (C, γ), onde C é a constante de regularização e γ é o parâmetro do kernel gaussiano. O GQ foi concebido com base em Keerthi e Lin (2003) que, a partir do comportamento assintótico do SVC com kernel gaussiano, identificaram características padrões do espaço de busca de (C, γ). Conforme evidenciado pela FIGURA 10, esse espaço é constituído por uma boa região de parâmetros e por áreas de *underfitting* e *overfitting*, separadas por uma curva de erro de generalização.

No capítulo 2, mostrou-se que a tradicional busca por *grid* (BG), ao pesquisar o espaço de parâmetros (C, γ), avalia todos os pontos da sua malha para encontrar a solução. Contudo, constata-se que, devido ao espaço de parâmetros ser formado por zonas de *underfitting* e *overfitting,* muitos dos pares investigados pela BG não são interessantes para o algoritmo SVC. Isso pode ser averiguado por meio da FIGURA 29, que ilustra a execução da BG em uma malha fictícia 9 x 9, onde cada ponto do *grid* equivale a uma combinação examinada de (C, γ).

FIGURA 29 - EXEMPLO DE EXECUÇÃO DA BUSCA POR *GRID*



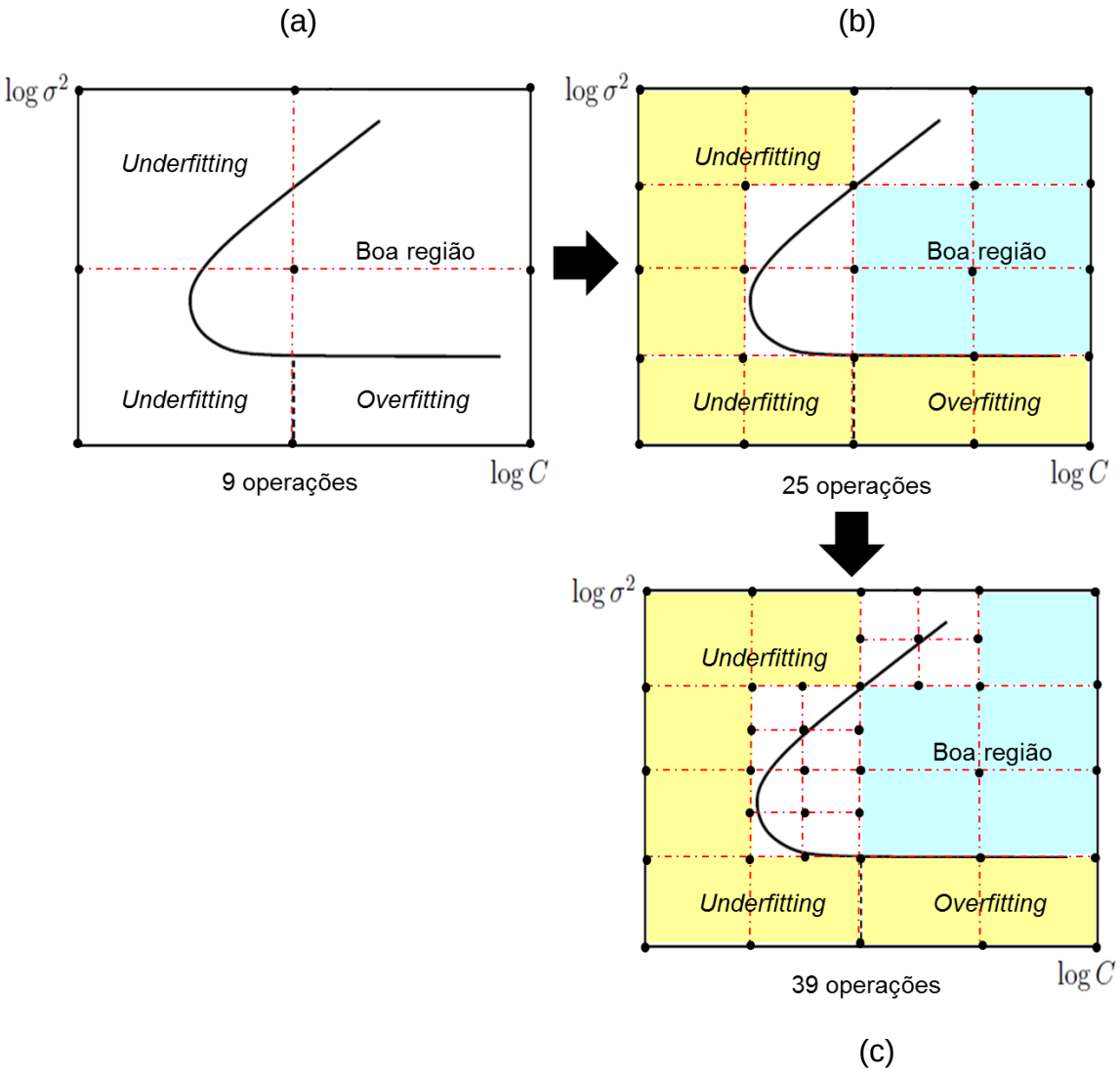
FONTE: A autora (2016).

Para o exemplo da FIGURA 29, repara-se que, dos 81 pontos analisados pela BG, cerca de 50% são externos à boa região. Logo, o tempo despendido nessas avaliações torna-se um desperdício, já que os mesmos não propiciam bons resultados ao SVC.

A partir dessas constatações, surgiu a ideia de combinar a técnica *quadtree* ao *grid,* o que deu origem ao método GQ. A finalidade da proposta é empregar a *quadtree* para identificar no espaço de busca de parâmetros, evidenciado por Keerthi e Lin (2003), as regiões de (C, γ) com características semelhantes (homogêneas) e desconsiderá-las da investigação.

Em outras palavras, uma vez que a *quadtree* divide apenas quadrantes heterogêneos, o fundamento do GQ consiste em reduzir a quantidade de operações efetuadas pelo *grid,* à medida que várias regiões do espaço deixam de ser avaliadas. Por exemplo: as zonas de *underfitting* e *overfitting*, por não serem úteis ao SVC, são entendidas pela *quadtree* como nós BRANCOS (externos à região de interesse) e, portanto, não se dividem. O mesmo ocorre com os quadrantes internos à boa região que, por estarem completamente inseridos na área de interesse, são compreendidos como nós PRETOS e também dispensam divisões. Logo, a *quadtree* limita o número de pares de parâmetros analisados, descartando a varredura completa do *grid.* A FIGURA 30 ilustra, em três etapas, a ideia central do método GQ.

FIGURA 30 - EXEMPLO DE EXECUÇÃO DO MÉTODO *GRID-QUADTREE*



FONTE: A autora (2016).

Pela FIGURA 30a, ao analisar o primeiro passo do método GQ, verifica-se que o espaço de parâmetros (C, γ) foi fragmentado em quatro quadrantes heterogêneos, pois todos contêm regiões internas e externas à boa região. Desta forma, é necessário subdividi-los em quatro novos quadrantes, conforme mostra a FIGURA 30b. Feito isso, nota-se que apenas três dos quadrantes resultantes, localizados na região da curva de erro de generalização, permanecem heterogêneos. Os demais representados pelas cores amarelo e azul claro, que correspondem respectivamente aos nós BRANCOS e PRETOS, deixam ser avaliados. Assim, divide-se somente os três quadrantes CINZA, pertencentes à fronteira da região, segundo se observa na FIGURA 30c.

É importante esclarecer que no método GQ a técnica *quadtree* é aplicada sobre um *grid.* Logo, estipulou-se que as divisões dos nós heterogêneos só podem ser efetuadas enquanto as laterais dos quadrantes não atingirem o tamanho equivalente à espessura da malha. Por exemplo: na FIGURA 30, considera-se que o *grid* utilizado no método GQ tem a mesma dimensão 9 x 9 do da BG da FIGURA 29. Desta forma, mesmo que ainda existam quadrantes heterogêneos na FIGURA 30c, o processo de divisão da *quadtree* deve ser encerrado. Isso ocorre, pois tais quadrantes já alcançaram o tamanho mínimo permitido (espessura do *grid*), estabelecido como critério de parada[9] do método GQ.

Tal qual a BG da FIGURA 29, os pontos situados nas intersecções dos quadrantes da FIGURA 30 referem-se aos pares de parâmetros (C, γ) testados pelo método GQ. Repara-se, ao comparar as FIGURAS 29 e 30c, que para este exemplo fictício a quantidade de pontos avaliada pelo GQ é inferior ao da BG, atendendo ao que foi idealizado.

Nesta seção, apresentou-se de maneira geral a ideia do método GQ e as expectativas em relação a ele (redução de cálculos). Na continuação, explicam-se em detalhes os procedimentos adotados para sua concretização.

4.2 MEDIDAS DE DESEMPENHO DO MÉTODO *GRID – QUADTREE*

Para melhor compreender a implementação do método *grid-quadtree*, é necessário iniciar a sua explicação pelas suas medidas de desempenho e pelo seu mecanismo de avaliação de parâmetros (C, γ). Entretanto, como o GQ surgiu dos fundamentos da BG, é imprescindível que se faça um paralelo entre o método proposto e a técnica tradicional.

Conforme visto, na busca por *grid* cada ponto da malha corresponde a um par de parâmetros (C, γ) por ela avaliado. De acordo com o passo 2 do QUADRO 2, isso significa que para cada combinação (C, γ) realiza-se uma validação cruzada *k-fold* no conjunto de dados, o que equivale a executar o procedimento descrito no QUADRO 15.

[9] O critério de parada do método GQ será explicado em mais detalhes na sequência deste capítulo.

QUADRO 15 - VALIDAÇÃO CRUZADA APLICADA NA SELEÇÃO DE PARÂMETROS DO SVC

PROCEDIMENTO: VALIDAÇÃO CRUZADA *K-FOLD* APLICADA AO ALGORITMO SVC
1. **PARA** cada par (C, γ)
2. Dividir o conjunto de dados em *k* subconjuntos de mesmo tamanho
3. Ajustar o algoritmo SVC com os parâmetros (C, γ)
4. Treinar o SVC utilizando *k* – 1 subconjuntos
5. Testar o SVC empregando o subconjunto restante
6. Repetir por *k* vezes os passos 4 e 5 até que todos os subconjuntos sejam testados
7. Obter a taxa de validação cruzada (VC) e o número de vetores suporte (VS) referentes às *k* execuções do procedimento

FONTE: A autora (2016).

Observando o passo 7 do QUADRO 15, percebe-se que cada par de parâmetros tem, após a sua avaliação, uma taxa de validação cruzada (VC) e uma quantidade de vetores suporte (VS) a ele associados. Por esses valores, pode-se verificar a qualidade da combinação (C, γ) e o seu impacto no treinamento do SVC.

Embora a BG, em sua versão original, use apenas a taxa VC como critério de seleção de parâmetros (passo 3 do QUADRO 2), neste trabalho, tanto o *grid-quadtree* quanto a busca por *grid* (aqui programados) empregam, além da VC, o número de vetores suporte como medida de desempenho de (C, γ). A análise desse último é importante ao passo que uma alta quantidade de VS revela um possível *overfitting* do algoritmo.

Por essa razão, tal indicador torna-se indispensável ao GQ, já que a *quadtree* necessita identificar no espaço de busca de parâmetros as regiões de *overfitting* e *underfitting*. Assim, para que a comparação dos dois métodos fosse realizada pelos mesmos critérios, optou-se por incluir na BG a informação do número de VS.

Pelo QUADRO 15, é possível certificar que a busca por *grid* é uma técnica de alto custo computacional, pois para cada parâmetro (C, γ) ela efetua *k* treinamentos no SVC. Desta forma, é razoável de imaginar que a comparação da BG com o GQ, além da taxa VC e quantidade de VS, deva ser feita pelo critério de tempo.

Contudo, uma vez que o tempo gasto por um método é influenciado pela sua forma de programação (estrutura e quantidade de comandos), não se achou justo confrontar duas técnicas de composições diferentes calculando-se o seu tempo computacional. Logo, decidiu-se por compará-los por meio da quantidade de operações efetuadas, visto que ambos empregam o mesmo princípio de análise de parâmetros (validação cruzada). Assim, nesta tese, uma operação é definida por:

Definição de operação: Uma operação corresponde à uma avaliação de parâmetros (C, γ) realizada pelos métodos GQ ou BG, que equivale a executar para o par (C, γ) em questão todo o procedimento descrito no QUADRO 15.

Nesta pesquisa, todos os cálculos relativos ao SVC (validação cruzada, treinamento, teste...), realizados pela BG e pelo o GQ, foram executados por meio de pacotes computacionais disponibilizados na biblioteca LIBSVM – *Library for Support Vector Machines* (CHANG; LIN, 2011). Esses pacotes, dentre os quais destacam-se o *svm-scale, svm-train* e *svm-predict*, cujas finalidades são respectivamente normalizar os dados, treinar e testar o SVC, foram utilizados em conjunto com a linguagem de programação VB.net. Nos códigos desenvolvidos, os comandos implementados em VB.net foram os responsáveis por executar os pacotes supracitados.

Desta forma, o conceito de operação, anteriormente apresentado, também pode ser estendido à aplicação do LIBSVM. Em outras palavras, realizar uma operação da BG ou da GQ equivale a chamar e a executar uma vez o aplicativo *svm-train,* com a opção de validação cruzada, do LIBSVM.

Por fim, para ilustrar a definição de operação, considera-se novamente o exemplo das FIGURAS 29 e 30, da onde se observa que a BG, por percorrer todo o espaço de busca, realizou 81 operações enquanto que a GQ efetivou apenas 39. Todavia, apesar de optar-se por computar operações ao invés de tempo, fica evidente que a quantidade de operações é proporcional ao tempo despendido. Logo, indiretamente, esse indicador também é mensurado.

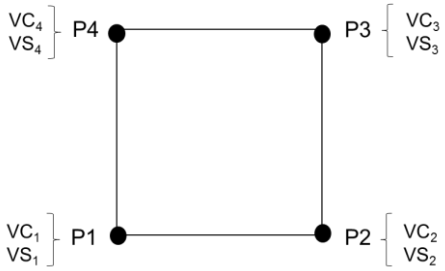
Em resumo, o principal critério empregado para medir a eficiência do método *grid-quadtree* é o número de operações por ele realizadas. Porém, juntamente com esse indicador, avalia-se a qualidade do parâmetro (C, γ) selecionado, cujo desempenho é dado pela taxa de validação cruzada (VC) e número de vetores suporte (VS) a ele vinculados.

4.3 CRITÉRIO DE DIVISÃO DOS QUADRANTES

Para que o método GQ possa funcionar, a técnica *quadtree* precisa identificar quando um quadrante é homo ou heterogêneo. Dessa maneira, necessita-se estabelecer critérios de referência para nortear a divisão ou não dos quadrantes.

Segundo o que mostra a FIGURA 31, no problema de seleção de parâmetros, cada nó (ou quadrante) tem quatro pontos (C, γ) a serem analisados pela *quadtree.* Esses pontos, conforme visto, possuem duas medidas muito importantes a eles associadas (VC e VS), provenientes da validação cruzada *k-fold*, que definem a condição de homogeneidade ou não de um quadrante.

FIGURA 31 - EXEMPLO DE NÓ OU QUADRANTE



FONTE: A autora (2016).

É com base nos valores de taxa VC e na quantidade de VS que a *quadtree* entende se um ponto pertence à boa região ou não. Portanto, para que ela distinga entre essas duas situações, é preciso delimitar antes do início do problema as características desejadas para a boa região.

Suponha que se queira, para um determinado conjunto de dados, que o método GQ encontre apenas pares de parâmetros que confiram ao SVC taxas de validação cruzada maiores ou iguais a 70% (VC ≥ 70%) e quantidades de vetores suporte menores ou iguais a 50% (VS ≤ 50%). Estabelecer essa vontade corresponde a restringir a boa região aos pares de parâmetros que atendam simultaneamente à essas duas condições.

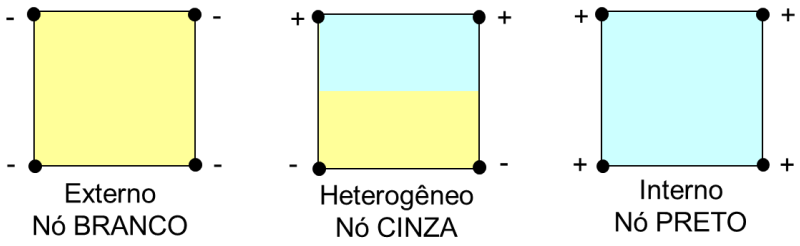
Em outras palavras, isso significa que a *quadtree* utilizará, para esse exemplo, os valores VC ≥ 70% e VS ≤ 50% como referência para classificar os pontos (C, γ). Assim, qualquer combinação de parâmetros que possua VC inferior a 70% e/ou VS superior a 50% é compreendida pela técnica como um ponto externo. Caso contrário, interno.

Neste trabalho, com o intuito de guiar a divisão dos quadrantes, convencionou-se uma regra de sinais baseada na classificação dos pontos, representada por (50):

$$Sinal = \begin{cases} +, & se\ P_i = (C_i, \gamma_i) \in\ boa\ região \\ -, & caso\ contrário \end{cases} \quad (50)$$

De acordo com (50), se o ponto for interno à boa região associa-se a ele o sinal positivo, caso contrário negativo. Assim, para o exemplo anterior, um ponto de VC = 80% e VS = 22% seria positivo “+” e outro de VC = 69% e VS=60% seria negativo “-“. Desta forma, dependendo dos sinais dos seus vértices, os quadrantes são categorizados pela *quadtree* como homogêneos ou heterogêneos. A FIGURA 32 ilustra algumas dessas possibilidades.

FIGURA 32 - CLASSIFICAÇÕES DE UM QUADRANTE



FONTE: A autora (2016).

Pela FIGURA 32, repara-se que se os quatro vértices do quadrante tiverem o mesmo sinal (todos positivos ou negativos) o nó será homogêneo e, consequentemente, não se dividirá. Nesse caso, segundo a terminologia da *quadtree*, se todos os sinais forem negativos trata-se de um nó BRANCO (externo à boa região), enquanto que todos positivos caracteriza um nó PRETO (interno à boa região).

Ainda pela FIGURA 32, verifica-se que os nós CINZAS, os heterogêneos, serão aqueles que possuírem tanto vértices negativos quanto positivos. Apesar da FIGURA 32 mostrar um exemplo de quadrante heterogêneo com dois vértices “+” e dois “–“, existem outras possibilidades de heterogeneidade. Assim, basta que apenas um ponto se diferencie dos demais, em termos de sinal, para que o quadrante seja dividido.

4.4 BALANCEAMENTO

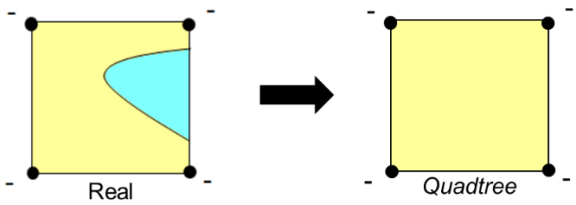
No capítulo 3, explicou-se que para aprimorar a convergência de um problema envolvendo *quadtree* é desejável que a mesma esteja balanceada. Isso significa que

a maior diferença entre os níveis de nós adjacentes não pode exceder a 1 para os vizinhos de aresta e a 2 para os de vértice.

No problema de seleção de parâmetros do SVC, será mostrado que realizar o balanceamento da *quadtree* é indispensável para que a técnica avalie todo o espaço de busca de (C, γ). Caso contrário, muitas áreas que contêm bons parâmetros deixam de ser analisadas e a reposta do método GQ torna-se insatisfatória. Isso ocorre em virtude do GQ referenciar-se somente nos vértices dos quadrantes para dividi-los ou não.

Conforme visto na FIGURA 32, quando um nó tem os seus quatro sinais negativos, a *quadtree* o interpreta como externo e negligencia a sua divisão. Contudo, existe uma situação em que o “bico” da curva, evidenciada por Keerthi e Lin (2003), intercepta um quadrante “negativo” pela sua lateral sem, no entanto, abranger seus vértices e afetar seus sinais. Quando isso acontece, o método GQ descarta tal nó, pois o entende como BRANCO, e deixa de averiguar os pares de parâmetros nele contidos. A FIGURA 33 ilustra esse equívoco, onde a parte esquerda retrata a situação real e a direita a compreendida pela *quadtree*.

FIGURA 33 - NÓ CINZA ENTENDIDO PELA *QUADTREE* COMO BRANCO

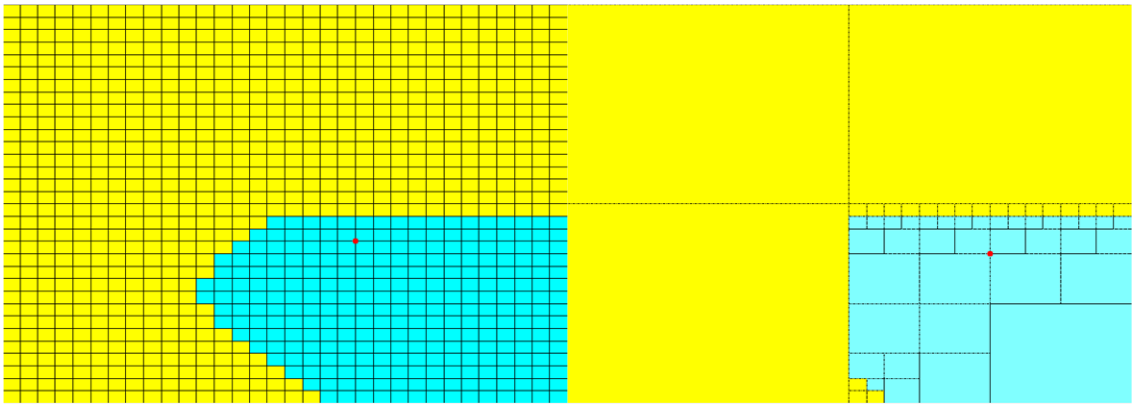


FONTE: A autora (2016).

Para melhor visualizar os efeitos da FIGURA 33 na resposta do problema, compara-se, por meio da FIGURA 34, as soluções gráficas da BG (à esquerda) com a do método GQ (à direita), quando esse emprega uma *quadtree* não balanceada. Essas soluções referem-se à busca de parâmetros (C, γ) para o conjunto de dados *Ionosphere*[10].

[10] As características dessa base de dados, disponibilizada no repositório LIBSVM, serão apresentadas no decorrer deste capítulo.

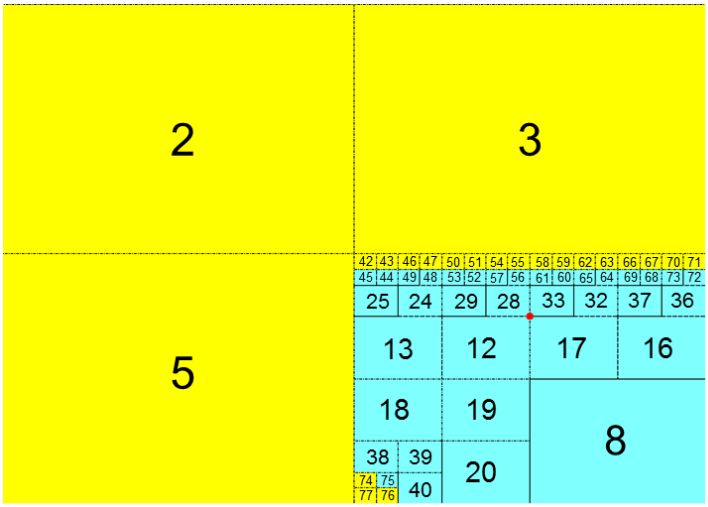
FIGURA 34 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DA BG E GQ SEM BALANCEAMENTO PARA O CONJUNTO *IONOSPHERE*



FONTE: A autora (2016).

Por tratarem-se da mesma base de dados, as soluções gráficas mostradas na FIGURA 34 deveriam ser semelhantes ou aproximadamente iguais. Todavia, a discrepância entre elas é bastante clara. A justificativa para tal se faz com base na FIGURA 35, que realça a numeração dos nós folhas da solução do GQ, apresentada na FIGURA 34.

FIGURA 35 - SOLUÇÃO GRÁFICA NUMERADA DO *GRID-QUADTREE* SEM BALANCEAMENTO



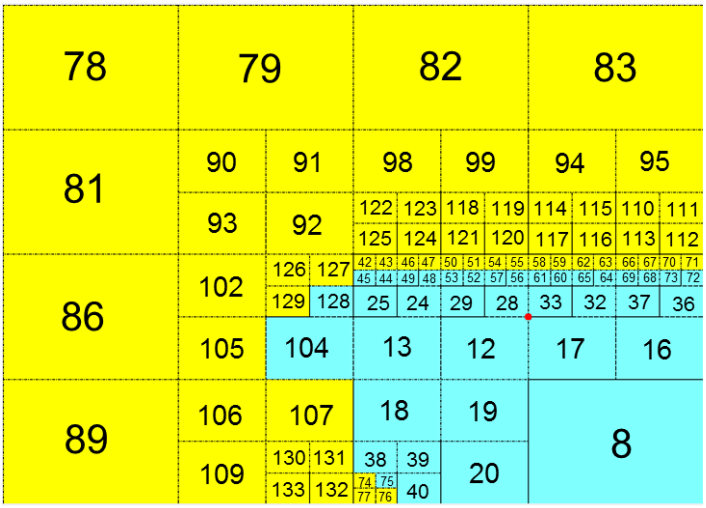
FONTE: A autora (2016).

Observando as FIGURAS 34 e 35, fica evidente que o quadrante 5 do GQ é um exemplo típico do que foi destacado na FIGURA 33 e a razão pela qual a *quadtree* não foi capaz de identificar corretamente o espaço de busca de (C, γ). Diante desse fato, deve-se pensar no balanceamento como uma maneira de contornar tal problema.

Ao analisar a FIGURA 35, nota-se que a *quadtree* está completamente desbalanceada, pois existem diversos nós, vizinhos de aresta e de vértice, que possuem diferenças de nível acima de 1 para o primeiro e de 2 para o segundo.

Neste caso, se o algoritmo do balanceamento (QUADRO 6) fosse aplicado na *quadtree* da FIGURA 35, seria fácil de identificar que os quadrantes 2, 3 e 5 seriam divididos logo no início da sua execução, o primeiro por critério de vértice e os demais por aresta. Ainda, tem-se que outros nós, além desses, seriam fracionados na continuação do processo. A FIGURA 36 ilustra o resultado final do balanceamento para a *quadtree* da FIGURA 35.

FIGURA 36 - SOLUÇÃO GRÁFICA DO GQ COM BALANCEAMENTO



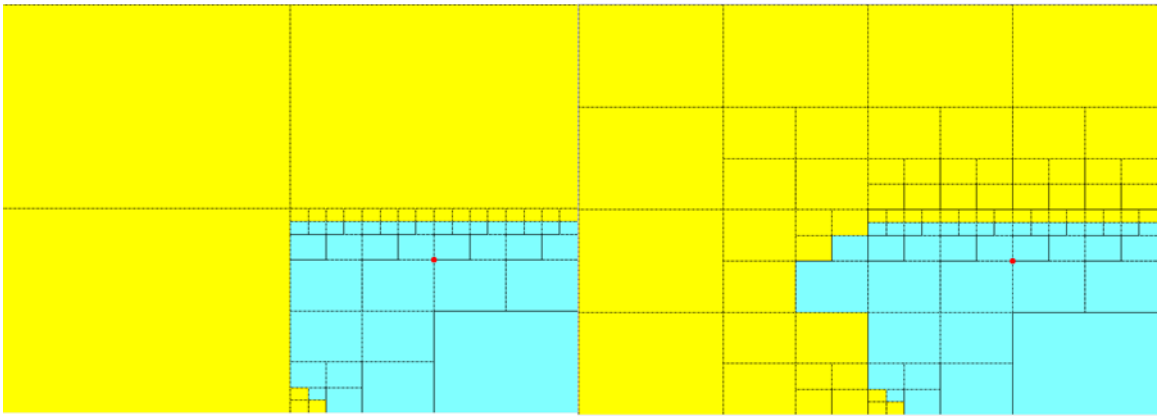
FONTE: A autora (2016).

Pela FIGURA 36, constata-se que o balanceamento é de certa forma uma alternativa para superar a dificuldade evidenciada nas FIGURAS 33 e 35, pois ele força a divisão de quadrantes supostamente homogêneos. Contudo, ao comparar a solução gráfica da FIGURA 36 com a da busca por *grid* (FIGURA 34 à esquerda), repara-se que a resposta da GQ, ainda assim, é diferente da fornecida pela técnica tradicional. Logo, conclui-se que mesmo que o balanceamento melhore a convergência do problema, para a seleção de parâmetros do SVC, somente a sua execução não é suficiente para encontrar uma boa solução. Desta maneira, para o bom funcionamento do método GQ, nesta tese, foi necessário desenvolver um procedimento denominado “refinamento”.

4.5 REFINAMENTO

A etapa do balanceamento, conforme visto no capítulo 3, evita mudanças bruscas entre os elementos da malha uma vez que limita a variação de tamanho de quadrantes vizinhos. Para o problema aqui considerado, esse procedimento é vantajoso, pois impede a presença de quadrantes grandes (pouco ou não avaliados) próximos a áreas de interesse, prevenindo a ocorrência da situação descrita pela FIGURA 33. Pela FIGURA 37, que coloca lado a lado as soluções do GQ sem e com balanceamento para o conjunto *Ionosphere*, observa-se que o balanceamento ocorre principalmente nos nós próximos à fronteira formada pela boa região e pelas as áreas de *underfitting* e *overfitting.*

FIGURA 37 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DO GQ SEM E COM BALANCEAMENTO



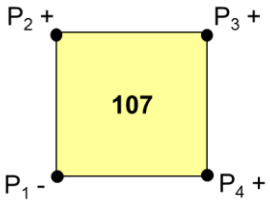
FONTE: A autora (2016).

Contudo, ao examinar em detalhes a FIGURA 37, nota-se que o balanceamento apenas equilibra a disposição e tamanho dos elementos da malha, sem, no entanto, avaliar as características dos quadrantes gerados. Desta forma, a sua finalidade é dividir os quadrantes para nivelar a malha e não analisar seus pontos em relação a taxa de validação cruzada (VC) e número de vetores suporte (VS).

Logo, pela FIGURA 36, evidencia-se que após o término do balanceamento existem quadrantes, como por exemplo o 107, que apesar de atender a todos os critérios de diferença de nível em relação a seus vizinhos, permanecem heterogêneos e de tamanho superior à espessura mínima da malha. Portanto, esses nós, em conformidade aos pressupostos do método GQ, deveriam continuar a ser investigados (divididos). Porém, como esse não é o propósito do balanceamento, o processo de

busca de parâmetros fica, até então, incompleto e a solução gráfica do GQ diferente da fornecida pela BG. A FIGURA 38 destaca os sinais e a heterogeneidade do nó 107.

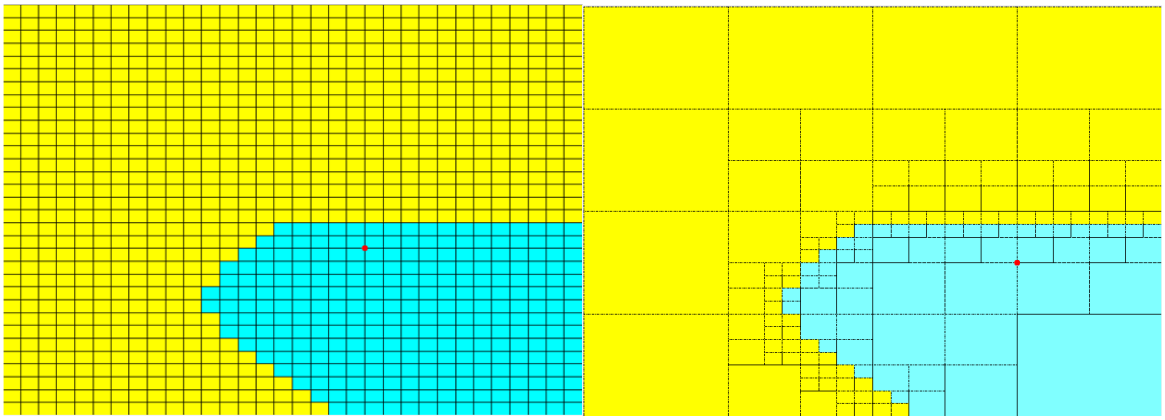
FIGURA 38 - DESTAQUE DOS SINAIS DO QUADRANTE 107



FONTE: A autora (2016).

Pela FIGURA 36, constata-se então a necessidade de se fazer uma análise mais criteriosa dos novos nós gerados. Considerando que esses novos quadrantes se concentram na fronteira da região de interesse, continuar a sua avaliação significa refinar a curva identificada por Keerthi e Lin (2003). Sendo assim, a etapa do refinamento detalha o formato e as características da curva, inicialmente delimitada pelo balanceamento. A FIGURA 39 compara a solução gráfica da BG (à esquerda) e do GQ após o refinamento (à direita).

FIGURA 39 - COMPARAÇÃO DAS SOLUÇÕES GRÁFICAS DA BG E GQ APÓS REFINAMENTO



FONTE: A autora (2016).

Pela FIGURA 39, verifica-se que a solução gráfica do *grid-quadtree* após refinamento, em termos de demarcação da boa região, é muito semelhante à da busca por *grid.* Desta forma, para que o GQ identifique corretamente o espaço de busca de parâmetros (C, γ) é necessário realizar tanto o balanceamento quanto o refinamento da *quadtree*. No QUADRO 16 expõe-se o pseudocódigo do procedimento de refinamento, desenvolvido nesta tese.

QUADRO 16 - PSEUDOCÓDIGO DO ALGORITMO DE REFINAMENTO DA *QUADTREE*

PROCEDIMENTO: REFINAMENTO
1. Inserir os nós gerados no balanceamento, NO_i, em uma lista R
2. **ENQUANTO** R não for vazia
3. **PARA** cada NO_i
4. Conduzir o procedimento de validação cruzada 5-*fold*, conforme o quadro 15
5. Determinar os sinais dos vértices (C, γ), segundo a relação (50)
6. Avaliar se o NO_i pode ser dividido, considerando os critérios de parada[11]
7. **SE** NO_i dividir
8. Inserir os quatro filhos de NO_i em R

FONTE: A autora (2016).

Em resumo, o objetivo do refinamento é avaliar se os nós criados após o balanceamento precisam ser ou não divididos. Conforme explicado, essa análise se faz pelos sinais dos vértices dos quadrantes, assinalados em função dos seus valores de VC e VS. No refinamento, à medida que os novos nós são fracionados os seus filhos são inseridos em uma lista R para posterior averiguação. Assim, dá-se continuidade à *quadtree* até que todos os quadrantes se tornem homogêneos ou atinjam ao critério de parada.

4.6 O MÉTODO *GRID-QUADTREE*

O intuito desta seção é descrever as particularidades do método *grid-quadtree*, contemplando desde as suas características gerais até a sua implementação. Primeiramente, explica-se a sua forma de inicialização e, na sequência, apresenta-se o estudo estatístico realizado para embasar a escolha da quantidade de pontos iniciais do método. Por fim, expõem-se os seus critérios de parada e o seu pseudocódigo.

4.6.1 Características gerais

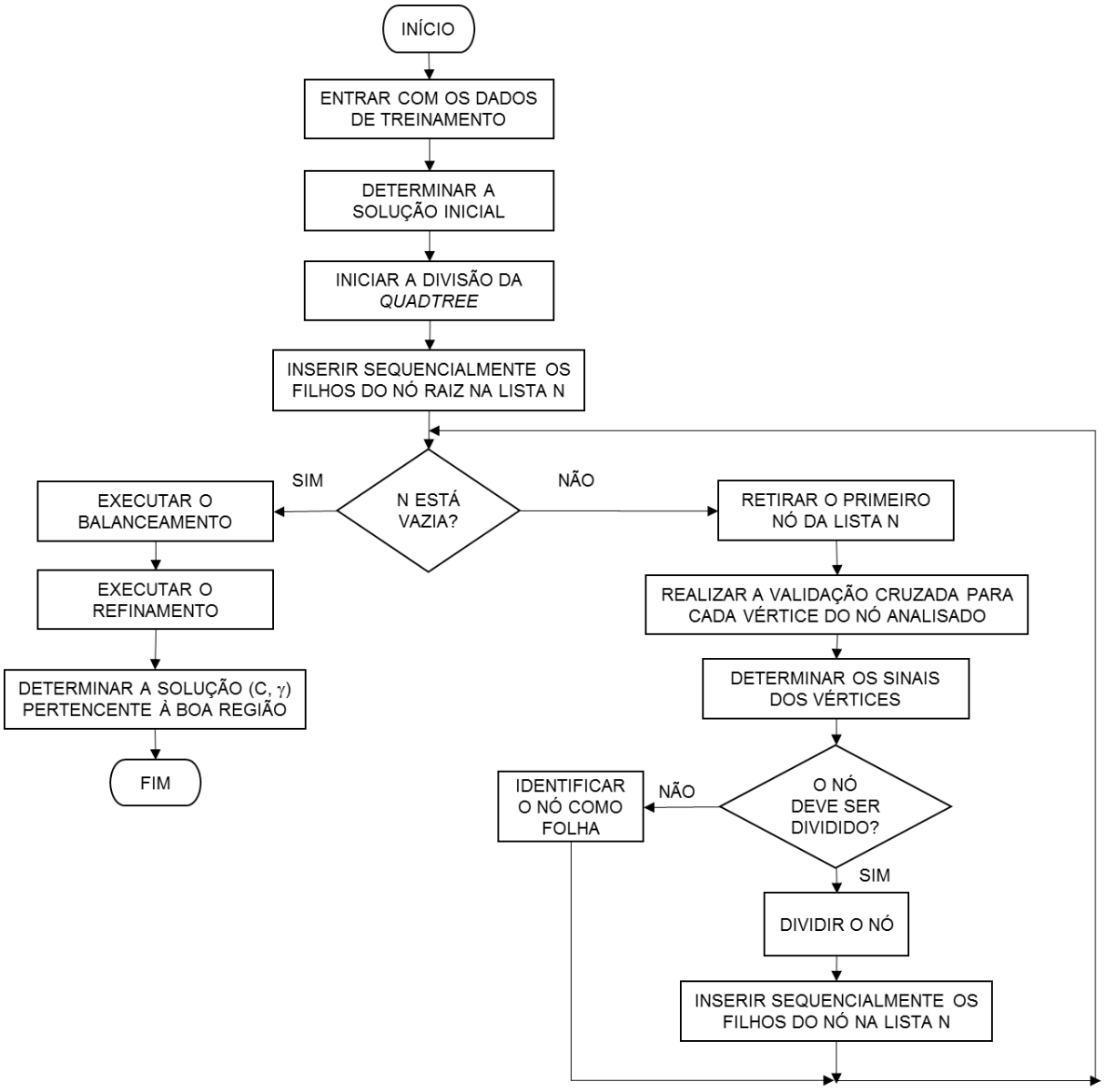
No método GQ, a procura dos parâmetros (C, γ) do SVC foi efetuada numa malha de tamanho 33 x 33, com os eixos de C e γ variando no intervalo de 2^{-8}, $2^{-7,5}$, 2^{-7},…, 2^{7}, $2^{7,5}$, 2^{8}. A definição dessas dimensões se sucedeu com base nos *grids*

[11] Serão explicados em detalhes na seção 4.6.5

empregados por Akay (2009), Hsu, Chang e Lin (2010), Keerthi e Lin (2003) e Pang *et al.* (2011), com a intenção de abranger as recomendações dadas por eles. Já o processo de validação cruzada *k-fold*, descrito no QUADRO 15, foi executado pelo GQ com parâmetro *k=5*, conforme sugere Kapp, Sabourin e Maupin (2012), que afirmam que esse valor é o mais utilizado para esse procedimento.

O fluxograma ilustrado na FIGURA 40 descreve de maneira geral o funcionamento do método GQ. As etapas de determinação da solução inicial e de encerramento do processo de divisão da *quadtree*, pertinentes à FIGURA 40, serão oportunamente explicadas nas próximas subseções.

FIGURA 40 – FLUXOGRAMA DO PROCESSO DE FUNCIONAMENTO DO MÉTODO GQ



FONTE: A autora (2016).

4.6.2 Inicialização do método

Dado que o método *grid-quadtree* não examina toda a malha 33 x 33 e o comportamento dos parâmetros, pertencentes ao espaço de busca, é inicialmente desconhecido, necessita-se estabelecer uma referência para a *quadtree*. Essa referência, que será determinada por um ponto, tem duas finalidades: orientar a divisão dos quadrantes (isto é, estipular as características da boa região) e indicar onde será realizada a primeira divisão da *quadtree*.

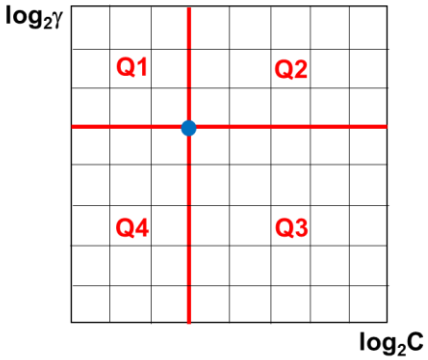
Na seção 4.3, que tratou dos critérios de divisão dos quadrantes, explicou-se que o método GQ requer dois valores padrões, um referente à taxa de validação cruzada (VC) e outro ao número de vetores suporte (VS), para identificar se um par de parâmetros (C, γ) pertence ou não à boa região. Esses valores padrões utilizados pelo GQ, conforme será mostrado, são oriundos do ponto de referência anteriormente citado.

A justificativa para adotar um ponto de referência para o GQ é que, de antemão, não se sabe como os dados irão se comportar com o kernel gaussiano. Ou seja, se for solicitado ao presente método que ele procure somente pares de parâmetros com taxa VC e quantidade de VS maiores ou menores à valores arbitrários, como por exemplo, VC ≥ 70% e VS ≤ 50%, é possível que o GQ se depare com um problema infactível. Em outras palavras, não há garantia que existam soluções (C, γ), que atendam simultaneamente aos dois critérios.

Para o exemplo anterior, se todo o espaço de busca fosse composto por parâmetros que concedessem ao SVC altas taxas VC (em torno de 95%) e número de VS superior a 51%, nenhuma solução seria encontrada pelo GQ, pois a segunda condição infringe ao que foi solicitado (VS ≤ 50%). Em contrapartida, a busca por *grid* não teria dificuldade em resolver esse mesmo problema, pois, uma vez que ela avalia ponto a ponto da malha, independe de qualquer referência.

Assim sendo, o método GQ precisa para o seu funcionamento de um ponto conhecido, que será usado como guia para todo o processo de divisão da *quadtree,* ou seja, é com base nos seus valores de VC e VS que os vértices dos quadrantes serão avaliados. Ainda, conforme ilustra a FIGURA 41 (ponto azul), é por meio dele que se dará a partição inicial do GQ.

FIGURA 41 - PONTO DE REFERÊNCIA E DIVISÃO INICIAL DO MÉTODO GQ



FONTE: A autora (2016).

Outro aspecto importante em relação ao ponto de referência é que ele garante para a *quadtree* a existência de pelo menos um vértice com sinal positivo[12]. Como normalmente, em função das propriedades do espaço de busca de (C, γ) o vértice superior esquerdo de Q1 (FIGURA 41) é negativo (zona de *underfitting*), essa diferença de sinais propicia a condição de heterogeneidade do quadrante e o início de todo o processo de busca da *quadtree*.

Entretanto, se ocorrer de todos os vértices de Q1 a Q4 (FIGURA 41) serem positivos[13], a *quadtree* não se fragmenta e o GQ considera o ponto azul como solução do problema. Desta forma, a escolha do ponto de referência, também entendido como solução inicial do GQ, deve ser feita de forma otimizada.

4.6.3 Determinação da solução inicial

A solução inicial (SI) do GQ deve ser apurada criteriosamente considerando um conjunto aleatório de pontos (C, γ). A sua seleção levando-se em conta um único ponto da malha poderia acarretar no *overfitting* do SVC. Um exemplo disso seria a escolha aleatória de um par (C, γ) cujo processo de validação cruzada *k-fold* resultasse nos valores VC = 95% e VS = 85%.

Observando essas medidas, constata-se que se trata de um ponto localizado na zona de *overfitting* do espaço de busca, visto que ele fornece alta taxa VC ao SVC, mas com grande quantidade de VS. Logo, se esses valores fossem usados para

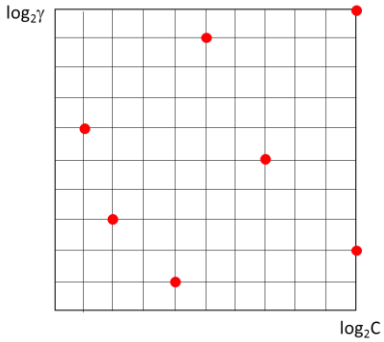
[12] Diz-se que o ponto azul da FIGURA 41 tem sinal positivo uma vez que ele determina as características da boa região e é considerado automaticamente pertencente a ela.

[13] Situações em que a boa região compreende todo o espaço delimitado pelos eixos de C e de γ.

referenciar o método GQ, esse provavelmente forneceria como resposta um par de parâmetros da região de *overfitting*, pois ele entenderia essa área como pertencente à boa região.

Desta forma, para escolher a SI do GQ deve-se considerar um número P de pontos gerados aleatoriamente e existentes sobre o *grid,* conforme mostra o exemplo da FIGURA 42. No procedimento adotado, analisa-se a quantidade média de vetores suporte fornecida pelos P pontos e, a partir de uma linha de corte, seleciona-se aquele de maior taxa de validação cruzada. O QUADRO 17 apresenta em detalhes esse processo.

FIGURA 42 - EXEMPLO DE PONTOS GERADOS ALEATORIAMENTE E PERTENCENTES À MALHA



FONTE: A autora (2016).

QUADRO 17 - PSEUDOCÓDIGO DO ALGORITMO DE DETERMINAÇÃO DA SOLUÇÃO INICIAL

PROCEDIMENTO: DETERMINAÇÃO DA SOLUÇÃO INICIAL DO GQ (PONTO DE REFERÊNCIA)

1. Gerar aleatoriamente P pontos (C, γ) pertencentes à malha 33 x 33
2. Para cada um dos P pares de parâmetros do passo 1, conduzir uma validação cruzada 5-*fold,* conforme quadro 15
3. Calcular a média aritmética do número de vetores suporte (VS), resultantes das P avaliações realizadas no passo 2
4. Desconsiderar todos os pares (C, γ) do passo 1, cujo procedimento do passo 2 resultou numa quantidade de VS superior à média calculada no passo 3.
5. Dentre os pontos não eliminados no passo 4, escolher como solução inicial do *grid-quadtree* aquele que fornece a maior taxa de classificação por validação cruzada (VC)

FONTE: A autora (2016).

Percebe-se, pelo QUADRO 17, que o intuito desse procedimento é referenciar a procura de parâmetros pela maior taxa VC possível sem, no entanto, considerar possibilidades que levem ao *overfitting* do SVC (alto quantidade de VS). Assim, o controle do número de VS pelo cálculo da média (linha de corte) é indispensável para

a exclusão de parâmetros ruins e, consequentemente, para a correta demarcação da boa região.

Em relação ao passo 2 do QUADRO 17, convém destacar que as P avaliações de validação cruzada 5-*fold,* realizadas para encontrar a SI, também são contabilizadas no cálculo de operações do método *grid-quadree*. Além disso, a escolha do valor de P influencia diretamente nessa quantidade total.

Alguns experimentos aleatórios, consolidados durante o desenvolvimento da tese, mostraram, por exemplo, que o GQ com P=5 e P=80 executou respectivamente 523 e 152 operações para encontrar a solução de parâmetros para o conjunto *Australian*[14]. Desta forma, motivados por esses testes iniciais, decidiu-se fazer uma análise estatística para determinar o valor adequado de P, a ser empregado na continuação desta pesquisa. O objetivo é estabelecer um P que não comprometa o desempenho do GQ, ou seja, que não agrave a quantidade de operações por ele efetuadas, e que forneça uma boa solução inicial.

4.6.4 Estudo estatístico para definir o número de pontos aleatórios P

Para estipular o valor de P, foram utilizadas seis bases de dados, dentre as quais cinco[15] são consideradas referência na área de classificação. No QUADRO 28 do apêndice, apresenta-se a relação dos trabalhos estudados na revisão de literatura desta tese, que analisaram esses mesmos dados. Diz-se que a *Fourclass* é a base de menor expressão, uma vez que somente Ho e Kleinberg (1996), dentre os autores do QUADRO 28, trabalharam com ela.

Essas bases, apesar de retiradas da biblioteca LIBSVM, são oriundas de outras coleções e pesquisas como, por exemplo, o repositório UCI (LICHMAN, 2013) e o projeto Statlog (MICHIE; SPIEGELHALTER, TAYLOR, 1994). A opção por obtê-las via LIBSVM se deu por, nessa biblioteca, elas já estarem disponibilizadas no formato exigido pelos seus próprios pacotes computacionais (*svm-train, svm-scale* e *svm-predict).* O QUADRO 18 descreve as características dessas bases.

[14] As características dessa base de dados, disponibilizada no repositório LIBSVM, serão apresentadas no decorrer deste capítulo.
[15] Provindas do repositório UCI e do Statlog.

QUADRO 18 - BASES DE DADOS EMPREGADAS NA DETERMINAÇÃO DE P

BASE DE DADOS	CLASSES	DADOS	ATRIBUTOS	NORMALIZAÇÃO	FONTE ORIGINAL
Sonar	2	208	60	[-1, 1]	UCI
Australian	2	690	14	[-1, 1]	Statlog
Fourclass	2	862	2	[-1, 1]	Ho e Kleinberg (1996)
German	2	1000	24	[-1, 1]	Statlog
Wine	3	178	13	[-1, 1]	UCI
Glass	6	214	9	[-1, 1]	UCI

FONTE: A autora (2016).

Em vista que o objetivo deste estudo é determinar um valor oficial para o parâmetro P a ser implantado no GQ, nesta etapa do trabalho não houve preocupação em dividir os dados do QUADRO 18 em conjuntos de treinamento e de teste. Ressalta-se que, neste momento, a finalidade não é averiguar a qualidade dos parâmetros (C, γ), obtidos pelo GQ e, sim, definir um P que permita encontrá-los de forma mais eficiente. Portanto, nesta avaliação, executou-se o procedimento do QUADRO 17 utilizando as bases de dados em sua totalidade.

Considerando que o ponto inicial do GQ é escolhido a partir da geração de P pontos aleatórios (passo 1 do QUADRO 17), o método GQ normalmente apresenta diferentes soluções a cada vez que é executado. Desta forma, para avaliar as variações nas respostas do GQ, em termos de número de operações, taxa de validação cruzada e quantidade de vetores suporte (medidas de desempenho definidas na seção 4.2), decidiu-se rodar o método 50 vezes e, para cada P verificado, mensurar seus intervalos de 95% de confiança (IC), suas médias e desvios padrões.

Dado que se percebeu, por meio de testes ocasionais, que o uso de P=5 aumentava consideravelmente o número de operações do GQ independentemente da base de dados investigada, optou-se por iniciar a avaliação de pontos pela quantidade P=20. Assim, neste estudo estatístico, foram examinados os valores P = 20, 40, 60, 80 e 100. As TABELAS 1 a 11 apresentam os resultados encontrados.

A TABELA 1, que aborda o percentual de operações realizadas pelo GQ em função de P, teve seus valores calculados com base na performance da busca por *grid*. A BG, se fosse empregada para resolver o mesmo problema aqui considerado, efetuaria para qualquer conjunto de dados 1089 operações, pois avaliaria todas as combinações de parâmetros existentes na malha 33 x 33. Então, em relação a esse valor, é que se determinaram os percentuais da TABELA 1.

TABELA 1 - IC DO NÚMERO DE OPERAÇÕES (%) REALIZADAS PELO GQ EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	19,0725 ± 1,6792	17,2360 ± 0,9578	19,1717 ± 0,9694	20,0955 ± 0,6679	**21,8898 ± 0,8154**
Australian	20,5326 ± 3,0539	19,6823 ± 3,1566	16,7530 ± 1,0021	17,6547 ± 0,9372	19,0119 ± 0,8780
Fourclass	19,5868 ± 1,2616	18,5014 ± 0,7965	19,5831 ± 0,5582	20,7052 ± 0,5071	**22,2424 ± 0,4780**
German	20,2314 ± 2,9464	21,3370 ± 2,4650	16,8430 ± 1,6120	18,6722 ± 1,4935	19,6180 ± 1,7092
Wine	16,7052 ± 2,3525	17,4031 ± 1,8603	16,7199 ± 1,2569	17,4160 ± 1,6175	19,5647 ± 1,1745
Glass	16,5216 ± 2,2823	15,0230 ± 1,2974	16,3893 ± 1,2086	17,9357 ± 1,0533	18,3214 ± 0,7687

FONTE: A autora (2016).

Para a escolha de P, dentre os três critérios estudados, deu-se maior relevância ao número de operações realizadas pelo GQ para encontrar a solução (C, γ). Assim, iniciou-se a análise estatística por tais resultados.

Observando apenas os valores médios da TABELA 1, também destacados na TABELA 2, repara-se que, para todos os conjuntos de dados, as menores quantidades médias de operações ocorreram para P= 20, 40 e 60. Contudo, verifica-se que os resultados mais expressivos foram os dos dois últimos, pois a única base em que P=20 se superou foi na *Wine*. Ainda no que se refere a esse conjunto, aponta-se que com P=20 o GQ realizou em média 16,7052% das operações da BG enquanto que com P=60 efetuou em média 16,7199%, que é um valor muito próximo ao do primeiro (diferença relativa de 0,09%), ressaltando mais uma vez o desempenho de P=60.

TABELA 2 - NÚMERO MÉDIO DE OPERAÇÕES (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	19,0725	**17,2360**	19,1717	20,0955	21,8898
Australian	20,5326	19,6823	**16,7530**	17,6547	19,0119
Fourclass	19,5868	**18,5014**	19,5831	20,7052	22,2424
German	20,2314	21,3370	**16,8430**	18,6722	19,6180
Wine	**16,7052**	17,4031	16,7199	17,4160	19,5647
Glass	16,5216	**15,0230**	16,3893	17,9357	18,3214

FONTE: A autora (2016).

Ao examinar os desvios padrões do percentual de operações executadas pelo GQ, constata-se pela TABELA 3 que os maiores valores se deram primeiramente para P=20 e em segundo lugar para P=40, razão pela qual os descarta como opção para P. Em contrapartida, novamente pela TABELA 3, nota-se que os menores desvios

padrões ocorreram para P=100, seguido de P=80. Consequentemente, os valores de P=60 foram intermediários aos dos demais.

TABELA 3 - DESVIO PADRÃO DO NÚMERO DE OPERAÇÕES (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	**6,0581**	3,4556	3,4973	2,4096	2,9417
Australian	11,0177	**11,3882**	3,6152	3,3813	3,1675
Fourclass	**4,5514**	2,8736	2,0140	1,8293	1,7247
German	**10,6299**	8,8931	5,8156	5,3883	6,1664
Wine	**8,4872**	6,7115	4,5347	5,8354	4,2373
Glass	**8,2340**	4,6805	4,3604	3,8000	2,7733

FONTE: A autora (2016).

Ainda no que diz respeito ao número de operações do GQ, percebe-se pela TABELA 1 que, apesar de muitos intervalos de confiança se sobreporem entre si, para as bases *Sonar* e *Fourclass,* os limites inferiores dos IC de P=100 foram maiores que os extremos superiores dos demais P, quando considerados os mesmos conjuntos de dados. Isso evidencia que o aumento na quantidade de pontos aleatórios pode impactar no total de operações efetuadas pelo GQ. Desta forma, julgou-se que 100 não é uma escolha viável para P.

Assim sendo, restou analisar somente as possibilidades P=60 e P=80, cuja decisão se embasou nos resultados de taxa de validação cruzada e quantidade de vetores suporte, provenientes dos pares (C, γ) encontrados pelo GQ. A finalidade consiste em avaliar se o valor de P influencia na capacidade do GQ em determinar parâmetros que forneçam altas taxas VC e baixo número de VS ao SVC. A TABELA 4 mostra os intervalos de 95% de confiança do percentual de validação cruzada em função de P.

TABELA 4 - IC DA TAXA DE VALIDAÇÃO CRUZADA (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	91,1346 ± 0,0858	90,9904 ± 0,1394	91,0577 ± 0,1234	91,0577 ± 0,0808	91,1539 ± 0,0660
Australian	86,3652 ± 0,0302	86,3565 ± 0,0372	86,3362 ± 0,0398	86,3768 ± 0,0344	86,3797 ± 0,0393
Fourclass	100,000 ± 0,000	100,000 ± 0,000	100,000 ± 0,000	100,000 ± 0,000	100,000 ± 0,000
German	76,8840 ± 0,0846	76,9540 ± 0,0424	76,8320 ± 0,0931	76,8980 ± 0,0586	76,9120 ± 0,0545
Wine	99,3034 ± 0,0973	99,3933 ± 0,0427	99,3708 ± 0,0600	99,3933 ± 0,0427	99,4045 ± 0,0374
Glass	72,1682 ± 0,2328	72,0280 ± 0,2158	72,3645 ± 0,2221	72,3832 ± 0,2255	72,4299 ± 0,2189

FONTE: A autora (2016).

Embora a TABELA 4 forneça os IC de todos os P investigados, o foco se dá aos valores 60 e 80. Analogamente ao que foi realizado para o percentual de operações do GQ, destaca-se na TABELA 5 as taxas médias da TABELA 4, para as duas opções em análise. Pela TABELA 5, observa-se que ambos os P permitiram ao GQ encontrar taxas médias de VC bastante satisfatórias. Todavia, os valores médios de VC fornecidos pelo GQ com P=80 foram superiores aos determinados por P=60, para quatro dos seis conjuntos avaliados. Nos dois restantes, ambos obtiveram o mesmo desempenho.

TABELA 5 - TAXA MÉDIA DE VC (%) PARA P=60 E P=80 E A RESPECTIVA VARIAÇÃO (%)

BASE DE DADOS	60	80	VARIAÇÃO = $(P_{60} – P_{80}) / P_{80}$
Sonar	**91,0577**	**91,0577**	0,0000
Australian	86,3362	**86,3768**	-0,0470
Fourclass	**100,000**	**100,000**	0,0000
German	76,8320	**76,8980**	-0,0858
Wine	99,3708	**99,3933**	-0,0226
Glass	72,3645	**72,3832**	-0,0258

FONTE: A autora (2016).

No que se refere aos desvios padrões da taxa VC, devido ao uso de P= 60 e P=80, evidencia-se que, para ambos os casos, esses foram relativamente pequenos. Porém, conforme se vê pela TABELA 6, os valores em função de P= 80 foram menores para as bases *Sonar, Australian, German* e *Wine.*

TABELA 6 - DESVIO PADRÃO DA TAXA DE VALIDAÇÃO CRUZADA (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	0,3096	0,5029	0,4451	**0,2914**	0,2379
Australian	0,1089	0,1342	0,1435	**0,1242**	0,1419
Fourclass	0,0000	0,0000	**0,0000**	**0,0000**	0,0000
German	0,3053	0,1528	0,3359	**0,2114**	0,1965
Wine	0,3509	0,1540	0,2165	**0,1540**	0,1348
Glass	0,8400	0,7786	**0,8011**	0,8135	0,7899

FONTE: A autora (2016).

Por fim, examinam-se os efeitos de P sobre a quantidade de vetores suporte. De maneira análoga aos demais critérios, expõem-se nas TABELAS 7, 8 e 9 respectivamente os IC da quantidade de VS em função de P, o número médio de VS para os valores 60 e 80 e os desvios padrões de VS para todos os P.

TABELA 7 - IC DA QUANTIDADE DE VETORES SUPORTE (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	48,8654 ± 1,4748	46,9327 ± 1,7063	47,7019 ± 1,5858	46,6250 ± 1,5637	48,5000 ± 1,4807
Australian	26,6638 ± 1,0961	27,0725 ± 1,4987	28,4145 ± 2,0419	26,7710 ± 1,4768	27,5333 ± 1,7786
Fourclass	2,4965 ± 0,0186	2,4756 ± 0,0179	2,4664 ± 0,0142	2,4455 ± 0,0088	2,4548 ± 0,0119
German	43,4580 ± 0,4260	43,2840 ± 0,2278	43,5600 ± 0,3482	43,5600 ± 0,3177	43,5520 ± 0,2966
Wine	25,2697 ± 0,9519	25,9663 ± 0,6586	26,5281 ± 0,7453	26,6292 ± 0,6992	26,3371 ± 0,6498
Glass	58,0467 ± 1,2493	57,3738 ± 1,2160	58,1776 ± 1,2889	58,6822 ± 1,3648	59,3084 ± 1,3964

FONTE: A autora (2016).

Para o critério vetores suporte, novamente o P=80 mostrou-se superior ao 60. Quando considerada a quantidade média de VS, tem-se, pela TABELA 8, que P=80 forneceu menores quantidades de VS para três dos conjuntos avaliados. Apesar da maior discrepância ter ocorrido na base *Australian*, destaca-se que pares (C, γ) que gerem em média 28,4145% de VS, como o obtido por P=60, são bons parâmetros para o SVC, em termos dessa medida, pois não caracterizam possibilidade de *overfitting.* Já em relação aos seus desvios padrões, nota-se, pela TABELA 9, que P=80 apresentou menores desvios em cinco dos casos averiguados.

TABELA 8 - QUANTIDADE MÉDIA DE VS (%) PARA P=60 E P=80 E A RESPECTIVA VARIAÇÃO (%)

BASE DE DADOS	60	80	VARIAÇÃO = (P_{60} – P_{80}) / P_{80}
Sonar	47,7019	**46,6250**	2,3097
Australian	28,4145	**26,7710**	**6,1391**
Fourclass	2,4664	**2,4455**	0,8546
German	**43,5600**	**43,5600**	0,0000
Wine	**26,5281**	26,6292	-0,3797
Glass	**58,1776**	58,6822	-0,8599

FONTE: A autora (2016).

TABELA 9 - DESVIO PADRÃO DA QUANTIDADE DE VETORES SUPORTE (%) EM FUNÇÃO DE P

BASE DE DADOS	20	40	60	80	100
Sonar	5,3207	6,1560	5,7210	**5,6415**	5,3418
Australian	3,9546	5,4068	7,3666	**5,3280**	6,4166
Fourclass	0,0673	0,0647	0,0514	**0,0318**	0,0430
German	1,5368	0,8217	1,2562	**1,1461**	1,0702
Wine	3,4342	2,3762	2,6890	**2,5227**	2,3444
Glass	4,5072	4,3869	**4,6502**	4,9237	5,0380

FONTE: A autora (2016).

Se considerado o desempenho do GQ em função de VC e VS, a quantidade de pontos P=80 deveria ser a escolhida. Porém, no início desse estudo definiu-se o percentual de operações realizado pelo GQ como o critério de maior peso. Desta forma, calculou-se a variação percentual dos valores médios das medidas: número de operações, VC e VS obtidos por P=60 e P=80 e os apresentou respectivamente nas TABELAS 10, 5 e 8. Além disso, comparam-se as respostas dessas variações na TABELA 11.

TABELA 10 - VARIAÇÃO (%) DO NÚMERO MÉDIO DE OPERAÇÕES ENTRE P=60 E P=80

BASE DE DADOS	60	80	VARIAÇÃO (%) = $(P_{60} – P_{80}) / P_{80}$
Sonar	19,1717	20,0955	-4,5970
Australian	16,7530	17,6547	-5,1074
Fourclass	19,5831	20,7052	-5,4194
German	16,8430	18,6722	-9,7964
Wine	16,7199	17,4160	-3,9969
Glass	16,3893	17,9357	-8,6219

FONTE: A autora (2016).

TABELA 11 - VARIAÇÃO (%) PARA AS MEDIDAS DE DESEMPENHO ENTRE P=60 E P=80

BASE DE DADOS	VARIAÇÃO (%) = $(P_{60} – P_{80}) / P_{80}$		
	TAXA MÉDIA VC	QUANTIDADE MÉDIA DE VS	NÚMERO MÉDIO DE OPERAÇÕES
Sonar	0,0000	2,3097	-4,5970
Australian	-0,0470	6,1391	-5,1074
Fourclass	0,0000	0,8546	-5,4194
German	-0,0858	0,0000	-9,7964
Wine	-0,0226	-0,3797	-3,9969
Glass	-0,0258	-0,8599	-8,6219

FONTE: A autora (2016).

Constata-se, pela TABELA 11, que as variações percentuais referentes às operações executadas pelo GQ são muito mais significativas do que as relacionadas aos critérios VC e VS. Ou seja, o ganho na redução de operações com o uso de P=60 é muito maior do que a perda que ele apresenta em termos de taxa VC ou elevação de VS. A única variação insatisfatória de P=60 em relação à P=80 foi o aumento de 6,1391% de vetores suporte para o conjunto *Australian* (TABELA 8), o qual já se explicou não impactar negativamente no SVC.

Logo, conclui-se que P=60 é um valor adequado a ser empregado no método GQ e será, portanto, o utilizado em todos os experimentos desta tese.

4.6.5 Critérios de parada

Dois critérios de parada, relacionados às divisões dos quadrantes, foram estabelecidos para o método GQ. O primeiro deles, já explicado em seções anteriores, refere-se à condição de homogeneidade do nó. Conforme visto, se os quatro vértices dos quadrantes tiverem sinais iguais, o mesmo não se dividirá. Consequentemente, quando todos os nós se tornarem homogêneos, o processo de divisão da *quadtree* se encerra. Entretanto, adotar apenas esse critério como condição de parada do GQ é insuficiente para interromper a sua execução.

Isso ocorre devido à presença dos nós que delimitam a curva de erro de generalização, proposta por Keerthi e Lin (2003). Por estarem localizados na fronteira, que separa os nós internos dos externos, muitos deles são permanentemente heterogêneos, pois possuem vértices pertencentes à boa região e outros não. Assim, se a *quadtree* for interrompida baseando-se somente na condição de homogeneidade, o GQ dividirá infinitamente os nós da fronteira, de forma que esses atingirão tamanhos extremamente pequenos, e o GQ, por sua vez, efetuará mais operações que a própria busca por *grid.*

Para superar essa dificuldade, criou-se um segundo critério de parada, vinculado à resolução da *quadtree.* Lembra-se, da seção 3.2, que a resolução consiste no número de vezes que o processo de decomposição é realizado. Logo, quanto maior a resolução da técnica, mais divisões ela efetuará e menores serão os lados dos quadrantes gerados.

Para o método GQ, equipara-se a resolução da *quadtree* ao espaçamento mínimo da malha, que neste trabalho estipulou-se 0,5. Em outras palavras, permite-se a fragmentação de um quadrante enquanto as suas laterais forem maiores que essa medida. Desta maneira, ao limitar a resolução da *quadtree* à espessura do *grid,* impede-se que o GQ entre em um processo de *looping* infinito. Além disso, torna-se justa a comparação entre os dois métodos, uma vez que os elementos mínimos de cada malha terão aproximadamente a mesma dimensão.

Diz-se que as dimensões são aproximadas, já que é possível, ao final da *quadtree,* deparar-se com quadrantes contendo laterais menores que a espessura do

grid. Por exemplo, se durante o processo de divisão houver um quadrante heterogêneo com lateral 0,6, pelo princípio de funcionamento da *quadtree,* ele ainda será dividido e sua dimensão final será 0,3.

Assim, o GQ se encerrará diante de duas situações: quando todos os seus nós forem homogêneos ou quando atingirem tamanho menor ou igual ao permitido. Na FIGURA 39 (à direita), visualiza-se que os quadrantes situados na fronteira da boa região, após refinamento, encontram-se na dimensão mínima admitida, que, neste caso, é similar a espessura do *grid* verificada na FIGURA 39 à esquerda.

4.6.6 Pseudocódigo

Explicados todos os detalhes pertinentes às configurações e ao desenvolvimento do método *grid-quadtree*, é possível apresentar no QUADRO 19 o seu pseudocódigo. Repara-se, pelo QUADRO 19, que muitos dos passos do GQ remetem a procedimentos descritos no decorrer desta tese.

QUADRO 19 - PSEUDOCÓDIGO DO MÉTODO *GRID-QUADTREE*

MÉTODO: *GRID-QUADTREE*
1. Determinar a solução inicial do método executando o procedimento do quadro 17 com P=60
2. Adotar a taxa VC e o número de VS da solução inicial como referência para todo o processo de divisão da *quadtree*
3. Realizar a primeira divisão da *quadtree* considerando o par (C, γ), determinado no passo 1, como o centro do nó raiz, conforme mostra a figura 41.
4. Inserir os nós filhos da raiz, NO_i, gerados no passo 3, em uma lista N
5. **ENQUANTO** N não for vazia
6. **PARA** cada NO_i
7. Conduzir o procedimento de validação cruzada 5-*fold*, conforme o quadro 15
8. Estabelecer os sinais dos vértices (C, γ), segundo a relação (50)
9. **SE** os quatro sinais de NO_i forem iguais ou alguma das laterais de NO_i **≤** 0,5 **ENTÃO**
10. NO_i é folha
11. **SENÃO**
12. Dividir NO_i
13. Inserir os filhos de NO_i em N
14. Executar o balanceamento da *quadtree* por meio do procedimento descrito no quadro 6
15. Realizar o refinamento da *quadtree* por meio do procedimento descrito no quadro 16
16. **PARA** todos os pares (C, γ) pertences à boa região
17. SOLUÇÃO = (C, γ) de maior taxa VC

FONTE: A autora (2016).

No passo 17 do método GQ (QUADRO 19), se houver dois ou mais pares de parâmetros possuidores da maior taxa de validação cruzada, deve-se escolher, entre eles, aquele que apresentar menor número de vetores suporte. Caso o empate persista, após a adoção desse critério, seleciona-se aquele que tiver menor quantidade de vetores suporte *Bound* (VS *Bound*).

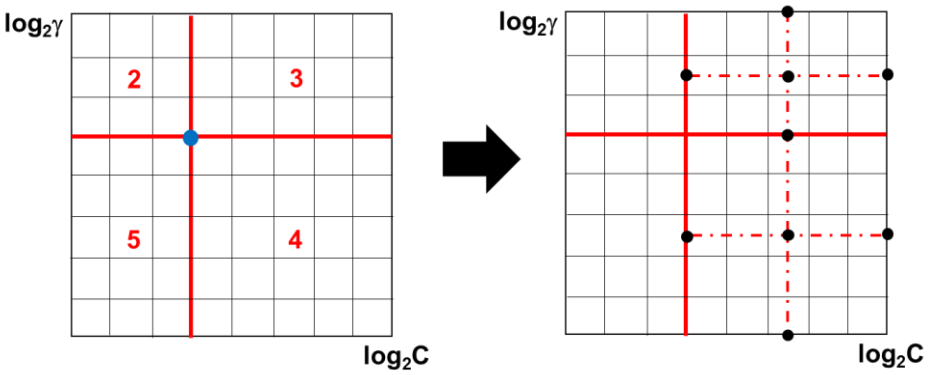
4.7 DIVISÃO UNIFORME E NÃO UNIFORME

É importante destacar que, no GQ, a divisão da *quadtree* raramente acontece de maneira regular, conforme ilustra a FIGURA 39. Na FIGURA 39, verifica-se que os quadrantes de mesmo nível, além de possuírem a mesma dimensão, tiveram seus vértices coincidindo com a demarcação do *grid.* Logo, nesse exemplo, os pontos examinados pelo GQ foram exatamente iguais a alguns dos averiguados pela BG.

Contudo, evidencia-se que a partição uniforme da *quadtree* somente ocorre se o ponto central da malha for o escolhido como solução inicial do GQ. Caso contrário, o método se depara com situações semelhantes à ilustrada na FIGURA 41, da onde percebe-se que os primeiros quadrantes gerados (nível 1 da árvore) não possuem a mesma dimensão.

Para melhor explicar a divisão irregular da *quadtree, a* FIGURA 43 simula o que aconteceria se o processo de fragmentação da FIGURA 41 fosse continuado. Nesse exemplo, considerou-se que apenas os filhos Q2 e Q3 da raiz (FIGURA 41), referentes aos nós 3 e 4 da árvore (FIGURA 43 à esquerda), precisavam ser divididos. Pela FIGURA 43 (à direita), repara-se que os novos pares (C, γ) avaliados pelo GQ, destacados em preto, não coincidem com os pontos do *grid*, pois localizam-se no interior de um elemento da malha, ou seja, entre duas posições consecutivas. Logo, essa situação permite ao GQ explorar regiões do espaço de busca onde a BG não tem acesso, o que pode acarretar no encontro de melhores soluções.

FIGURA 43 - EXEMPLO DE FRAGAMENTAÇÃO IRREGULAR



FONTE: A autora (2016).

Para finalizar, mostra-se pela FIGURA 43 que somente a primeira divisão da *quadtree* gera quadrantes irmãos de tamanhos distintos. Nas seguintes, pelas divisões serem efetuadas no ponto médio das laterais do nó avaliado, os quatro filhos serão iguais. Todavia, quando observados por nível, os quadrantes terão dimensões diferentes e, portanto, a solução gráfica do problema apresentará uma malha irregular.

Ainda, uma vez que o nó, em muitas situações, possui formato retangular, o critério de parada do GQ está vinculado separadamente às suas laterais. Isto é, se apenas um dos lados do quadrante possuir tamanho menor ou igual 0,5, já se encerra a sua divisão independentemente da dimensão da outra lateral.

4.8 VALIDAÇÃO DO MÉTODO *GRID-QUADTREE*

Com o intuito de validar o método *grid-quadtree*, comparou-se o seu desempenho com o da tradicional busca por *grid.* Essa técnica, além de motivar o desenvolvimento do GQ, é utilizada na validação de novos métodos de seleção de parâmetros do SVC. Os dados empregados nesta fase do trabalho, por receberem diferentes tratamentos, foram separados em dois grupos, conforme se explica adiante.

4.8.1 Características da busca por *grid*

A implementação computacional da busca por *grid* foi realizada, conforme o procedimento do QUADRO 2, empregando-se a linguagem VB.net. Tal qual ao GQ, todos os cálculos efetuados pela BG, relativos ao SVC, foram executados por meio dos pacotes computacionais do LIBSVM. Desta forma, por usarem os mesmos

mecanismos de cálculo (validação cruzada, treinamento e teste), qualquer discrepância entre os resultados dos métodos GQ e BG se dará unicamente pela determinação de diferentes parâmetros (C, γ) e não por influência dos seus meios de programação.

De forma análoga ao GQ, considerou-se para a busca por *grid* uma malha de tamanho 33 x 33, com os eixos de C e γ variando no intervalo de 2^{-8}, $2^{-7,5}$, 2^{-7},…, 2^{7}, $2^{7,5}$, 2^{8}. O processo de validação cruzada *k-fold* foi igualmente realizado com parâmetro *k=5*.

4.8.2 Dados estudados

Para validar o método GQ, utilizaram-se bases de dados referências na área de classificação, que foram novamente obtidas via repositório LIBSVM, com exceção da *Circle and Square*[16], que proveio de Carpenter, Grossberg e Reynolds (1991). Relembra-se, no entanto, que as bases retiradas do LIBSVM são todas originárias de outras fontes, conforme será propriamente indicado. Nos QUADROS 29 e 30 do apêndice, relacionam-se as bases de dados estudadas com os trabalhos apresentados na revisão de literatura.

Tendo em vista que muitas das bases aqui consideradas são disponibilizadas no todo, sem distinção entre seus conjuntos de treinamento e de teste, necessitou-se abordá-las de forma distinta. Portanto, dividiu-se os dados de estudo em dois grupos. O primeiro refere-se às bases cujos conjuntos de treinamento e de teste foram separados aleatoriamente e o segundo àquelas que já possuíam seus conjuntos bem definidos e assim divulgados. A seguir, detalha-se cada um dos grupos.

4.8.2.1 Grupo 1– Bases de dados separadas aleatoriamente

Neste grupo, abordam-se as bases de dados disponibilizadas sem diferenciação de conjuntos, cujas características originais estão descritas no QUADRO 20.

[16] Dentre as bases de dados estudadas neste trabalho, essa é a única que não faz parte da biblioteca LIBSVM.

QUADRO 20 - BASES DE DADOS PERTENCENTES AO GRUPO 1

BASE DE DADOS	CLASSES	DADOS	ATRIBUTOS	NORMALIZAÇÃO	FONTE ORIGINAL
Liver Disorders	2	345	6	[-1, 1]	UCI
Ionosphere	2	351	34	[-1, 1]	UCI
Breast Cancer	2	683	10	[-1, 1]	UCI
Diabetes	2	768	8	[-1, 1]	UCI
Circle and Square	2	1000	2	SEM	Carpenter, Grossberg e Reynolds (1991)
Mushroom	2	8124	112	[0, 1]	UCI
Iris	3	150	4	[-1, 1]	UCI
Svmguide 2	3	391	20	SEM	Hsu, Chang e Lin (2010)
Vehicle	4	846	18	[-1, 1]	Statlog
Segment	7	2310	19	[-1, 1]	Statlog

FONTE: A autora (2016).

Hsu, Chang e Lin (2010) recomendam que, antes de se utilizar o SVC, os dados sejam normalizados linearmente nos intervalos [-1, 1] ou [0, 1]. Segundo os autores, esse procedimento evita que os atributos de maior valor (pertencentes a grandes intervalos) dominem os de menor valor. Em concordância a essa orientação, normalizou-se as bases *Circle and Square* e *Svmguide* 2 no intervalo [-1, 1].

A separação das bases do QUADRO 20 foi realizada de forma aleatória, na composição: 80% dos dados para o conjunto de treinamento e 20% para o de teste. Tal divisão foi executada de maneira equilibrada, mantendo-se a proporção de cada classe para ambos os conjuntos. Visando descartar qualquer possibilidade de bons resultados devido a uma única partição aleatória dos dados, decidiu-se repetir esse procedimento por 30 vezes. Desta forma, geraram-se para cada base de dados trinta diferentes conjuntos de treinamento e de teste, denominados A1 a A30, cujas dimensões estão descritas no QUADRO 21.

QUADRO 21 - TAMANHO DOS CONJUNTOS DE DADOS GERADOS ALEATORIAMENTE

BASE DE DADOS	CONJUNTO	FINALIDADE	DADOS
Liver Disorders	A1... A30	Treinamento	276
		Teste	69
Ionosphere	A1... A30	Treinamento	281
		Teste	70
Breast Cancer	A1... A30	Treinamento	546
		Teste	137
Diabetes	A1... A30	Treinamento	614
		Teste	154
Circle and Square	A1... A30	Treinamento	800
		Teste	200
Mushroom	A1... A30	Treinamento	6499
		Teste	1625
Iris	A1... A30	Treinamento	120
		Teste	30
Svmguide 2	A1... A30	Treinamento	313
		Teste	78
Vehicle	A1... A30	Treinamento	677
		Teste	169
Segment	A1... A30	Treinamento	1848
		Teste	462

FONTE: A autora (2016).

Percebe-se pelo QUADRO 21 que, no total, o grupo 1 compreende trezentos conjuntos de treinamento e de teste, que serão empregados na avaliação de desempenho do GQ e BG.

4.8.2.2 Grupo 2– Bases de dados originalmente separadas

Diferentemente do grupo 1, o grupo 2 refere-se às bases de dados cujos conjuntos de treinamento e de teste já se encontram divulgados separadamente no LIBSVM. Para esses casos, manteve-se as suas divisões iniciais conforme mostra o QUADRO 22, que apresenta as características originais de cada base. É importante ressaltar que a conservação da divisão dos conjuntos propicia a reprodução dos resultados do GQ e a sua comparação com outros métodos, se assim desejado por pesquisadores da área.

QUADRO 22 - BASES DE DADOS PERTENCENTES AO GRUPO 2

BASE DE DADOS	CONJUNTO	CLASSES	DADOS	ATRIBUTOS	NORMALIZAÇÃO	FONTE ORIGINAL
A1A	Treinamento	2	1605	123	[0, 1]	UCI
	Teste		30956			
Splice	Treinamento	2	1000	60	[-1, 1]	UCI
	Teste		2175			
Svmguide 1	Treinamento	2	3089	4	SEM	Hsu, Chang e Lin (2010)
	Teste		4000			
Svmguide 3	Treinamento	2	1243	21	SEM	Hsu, Chang e Lin (2010)
	Teste		41			
W1A	Treinamento	2	2477	300	[0, 1]	Platt (1998)
	Teste		47272			
DNA	Treinamento	3	2000	180	[0, 1]	Statlog
	Teste		1186			
Satimage	Treinamento	6	4435	36	[-1, 1]	Statlog
	Teste		2000			
Svmguide 4	Treinamento	6	300	10	SEM	Hsu, Chang e Lin (2010)
	Teste		312			
Pendigits	Treinamento	10	7494	16	SEM	UCI
	Teste		3498			
Vowel	Treinamento	11	528	10	[-1, 1]	UCI
	Teste		462			

FONTE: A autora (2016).

Em conformidade a Hsu, Chang e Lin (2010), normalizou-se as bases *Svmguide1, Svmguide3* e *Svmguide4* no intervalo [-1, 1]. Já para a *Pendigits*, devido à muitos dos seus atributos originais serem iguais à zero, considerou-se o intervalo [0, 1], conforme sugestão do próprio pacote computacional *svm-scale* do LIBSVM. Com exceção dessas normalizações, as características das demais bases foram mantidas exatamente como descrito no QUADRO 22.

4.8.3 Comparação dos resultados

Para validar o método proposto, os resultados do GQ foram confrontados com os da BG levando-se em conta a qualidade da solução (C, γ) e a quantidade de operações necessárias para obtê-la. Entende-se por qualidade da solução a taxa VC e o número de VS vinculados ao par de parâmetros (C, γ) determinado.

Exclusivamente para as bases binárias, apresenta-se também a quantidade de vetores suporte *Bound*, que é a parcela de VS correspondente aos erros de classificação. Para as bases multiclasses esse indicador não é mostrado, pois os

relatórios de resposta de validação cruzada do LIBSVM não contabilizam o seu valor final.

Dado que o par (C, γ) é obtido pelo GQ e pela BG junto aos conjuntos de treinamento, para avaliar a eficiência dos parâmetros encontrados, esses foram empregados na avaliação dos conjuntos de teste. A partir dessa análise, compararam-se a acurácia e a quantidade de acertos do SVC, quando usados os parâmetros fornecidos por cada um dos métodos.

De maneira complementar, para algumas bases de dados, determinou-se a matriz de confusão do SVC. Essa matriz confronta as classificações preditas pelo algoritmo com os valores reais de cada classe. Assim, por meio dela, é possível avaliar a capacidade de discriminação do SVC.

QUADRO 23 - MATRIZ DE CONFUSÃO GENÉRICA

CLASSE	PREDITA C_1	PREDITA C_2	...	PREDITA C_K
VERDADEIRA C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_k)$
VERDADEIRA C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_2, C_k)$
⋮	⋮	⋮	⋱	⋮
VERDADEIRA C_K	$M(C_k, C_1)$	$M(C_k, C_2)$	...	$M(C_k, C_k)$

FONTE: A autora (2016).

O QUADRO 23 apresenta a forma genérica da matriz de confusão, onde os elementos da diagonal principal indicam os acertos em cada uma das classes e os demais $M(C_i, C_j)$ representam os erros na classificação. Já o QUADRO 24 resume o processo de validação do método GQ, explicado no decorrer desta subseção.

QUADRO 24 - RESUMO DAS ETAPAS DE VALIDAÇÃO DO MÉTODO *GRID-QUADTREE*

ETAPA	**MEDIDA DE DESEMPENHO COMPARADAS**
1) Determinação de parâmetros (C, γ) – CONJUNTO DE TREINAMENTO	• Número de operações executadas pelo método • Taxa de validação cruzada de (C, γ) • Quantidade de vetores suporte vinculada à (C, γ) • Parcela de VS *bound* associada à (C, γ) – (somente para as bases binárias)
2) Avaliação dos parâmetros (C, γ) – CONJUNTO DE TESTE	• Quantidade de acerto do SVC • Acurácia do SVC

FONTE: A autora (2016).

Por fim, relembra-se que o GQ, por iniciar a partir de pontos aleatórios, pode apresentar diferentes soluções a cada vez que é executado. Desta forma, apesar das variações entre os resultados serem pequenas, é possível que se obtenha melhores ou piores soluções a cada execução. Contudo, para que a comparação com a BG fosse justa, para cada conjunto de treinamento dos grupos 1 e 2, rodou-se o GQ apenas uma vez, independente se a resposta fornecida por ele seria ou não o seu melhor resultado. Esse procedimento, além de honesto, simula uma aplicação real do método GQ, pois os usuários do SVC, ao selecionar seus parâmetros, o aplicariam somente uma vez.

4.9 CARACTERÍSTICAS DO COMPUTADOR E VERSÕES DOS *SOFTWARES*

Os experimentos realizados nesta tese foram executados em um computador de configuração Intel (R) Core (TM) i7 – 3517U, 2,40 GHz com 4,00 GB de memória RAM, utilizando-se o *software Microsoft Visual Studio* 2012 (VB.net) em conjunto com os pacotes computacionais do LIBSVM, versão 3.18.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Este capítulo tem por finalidade apresentar os resultados computacionais obtidos pelo método *grid-quadtree* (GQ) e compará-los com os da busca por *grid* (BG). Primeiramente, avaliam-se os resultados encontrados para as bases de dados do grupo 1, as separadas aleatoriamente, e na sequência as do grupo 2, aquelas que já possuíam os seus conjuntos de treinamento e de teste devidamente divididos no LIBSVM.

5.1 RESULTADOS COMPUTACIONAIS DO GRUPO 1

Conforme visto no capítulo 4 (QUADRO 21), o grupo 1 de dados constitui-se de dez bases, separadas de forma aleatória e balanceada, que deram origem a 30 conjuntos de treinamento e de teste cada, denominados A1 a A30. A partir da análise desses conjuntos, sucederam 30 tabelas de resultados computacionais, sendo 3 por base. Em virtude da alta quantidade de respostas obtidas, muitas dessas tabelas serão apresentadas no apêndice desta tese.

Dentre as bases que compõem o grupo 1, a *Mushroom* é considerada a mais importante a ser analisada, devido ao seu grande número de dados (6499) e de características (112). Afirma-se isso uma vez que é um desafio encontrar o par de parâmetros (C, γ) em um conjunto de treinamento de tal dimensão, realizando operações de validação cruzada 5-*fold.* Logo, muitos são os cálculos efetuados na determinação de parâmetros da *Mushroom*, o que impacta diretamente no tempo computacional da procura. Em vista da relevância dos resultados alcançados e da expressividade da base *Mushroom*, optou-se por iniciar esta seção pela sua explicação.

De acordo com o exposto, cada base do grupo 1 possui três tabelas de respostas a ela vinculadas: duas que mostram os resultados obtidos por meio do conjunto de treinamento (características dos parâmetros e número de operações) e outra utilizando-se o conjunto de teste (acurácia e quantidade de acertos). No que diz respeito à *Mushroom,* a TABELA 12 é que a apresenta as características dos parâmetros ótimos (C, γ), encontrados pela BG e pelo GQ para os seus 30 conjuntos avaliados.

Ao interpretar a primeira e segunda linhas da TABELA 12, tem-se que essas indicam respectivamente, para o conjunto A1, as propriedades dor par (C, γ) estipulados pelas técnicas BG e GQ. Para A1, verifica-se que a BG determinou como solução ótima os parâmetros (181,0193; 0,0039) e o GQ (256,0000; 0,0039), ambos fornecidos com taxa de validação cruzada 100%. Em relação aos vetores suporte, as duas repostas estão associadas a 174 VS, que representam 2,6773% do total de dados de treinamento, sendo que nenhum deles é considerado erro de classificação (VS*Bound* = 0).

A análise das demais linhas da TABELA 12 é feita de maneira similar à do conjunto A1. Contudo, destacam-se que as situações em que aparecem o termo "BG/GQ" significam que, para certo conjunto de dados, ambos os métodos encontraram o mesmo par de parâmetros (C, γ) como resposta. Para a base *Mushroom,* observa-se que isso ocorreu em 21 dos casos estudados.

Por fim, pela TABELA 12, nota-se que a BG e o GQ determinaram, em todas as ocasiões, soluções com taxa VC igual a 100% e quantidades de vetores muito pequenas (inferiores ou iguais a 3,0005%). Relembra-se que o baixo percentual de VS, neste caso, descaracteriza a possibilidade de *overfitting* da função. Desta forma, na TABELA 12 constam parâmetros de alta qualidade, já que possuem alta taxa VC e baixo VS.

TABELA 12 - PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *MUSHROOM*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG	181,0193	0,0039	100	174	2,6773	0
	GQ	256,0000	0,0039	100	174	2,6773	0
A2	BG/GQ	45,2548	0,0039	100	178	2,7389	7
A3	BG/GQ	64,0000	0,0039	100	169	2,6004	2
A4	BG/GQ	90,5097	0,0039	100	173	2,6619	1
A5	BG/GQ	45,2548	0,0039	100	174	2,6773	7
A6	BG/GQ	64,0000	0,0039	100	179	2,7543	3
A7	BG/GQ	64,0000	0,0039	100	163	2,5081	3
A8	BG/GQ	45,2548	0,0039	100	161	2,4773	7
A9	BG/GQ	128,0000	0,0039	100	185	2,8466	1
A10	BG/GQ	64,0000	0,0039	100	170	2,6158	3
A11	BG	181,0193	0,0039	100	183	2,8158	0
	GQ	256,0000	0,0039	100	183	2,8158	0
A12	BG/GQ	64,0000	0,0039	100	183	2,8158	4

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A13	BG/GQ	45,2548	0,0039	100	174	2,6773	7
A14	BG	181,0193	0,0039	100	180	2,7697	0
	GQ	256,0000	0,0039	100	180	2,7697	0
A15	BG/GQ	45,2548	0,0039	100	169	2,6004	6
A16	BG/GQ	64,0000	0,0039	100	166	2,5542	2
A17	BG	45,2548	0,0039	100	180	2,7697	6
	GQ	48,2933	0,0039	100	181	2,7850	5
A18	BG/GQ	256,0000	0,0039	100	189	2,9081	0
A19	BG/GQ	128,0000	0,0039	100	195	3,0005	0
A20	BG/GQ	64,0000	0,0039	100	181	2,7850	3
A21	BG/GQ	128,0000	0,0039	100	184	2,8312	0
A22	BG/GQ	64,0000	0,0039	100	172	2,6466	3
A23	BG	64,0000	0,0039	100	161	2,4773	3
	GQ	69,7925	0,0039	100	161	2,4773	2
A24	BG	45,2548	0,0039	100	171	2,6312	6
	GQ	90,5097	0,0039	100	177	2,7235	1
A25	BG/GQ	64,0000	0,0039	100	175	2,6927	2
A26	BG	64,0000	0,0039	100	158	2,4311	3
	GQ	48,2933	0,0039	100	157	2,4158	5
A27	BG/GQ	45,2548	0,0039	100	168	2,5850	7
A28	BG	128,0000	0,0039	100	178	2,7389	0
	GQ	256,0000	0,0039	100	178	2,7389	0
A29	BG	45,2548	0,0039	100	181	2,7850	6
	GQ	32,0000	0,0039	100	187	2,8774	12
A30	BG/GQ	256,0000	0,0039	100	177	2,7235	0

FONTE: A autora (2016).

Na TABELA 13, compara-se o número de operações realizadas pelos métodos BG e GQ para encontrar os parâmetros da TABELA 12. Ressalta-se que a busca por *grid,* por calcular todas as combinações de (C, γ) no espaço de busca 33 x 33, efetuou para cada conjunto de dados 1089 operações. Com base nesse valor é que se avaliou o desempenho do método *grid-quadtree.*

Assim, para o conjunto A2, em que ambos os métodos indicaram a mesma solução, constata-se que enquanto a BG realizou 1089 operações para determinar o par (C, γ) o GQ executou somente 161, o que representa 14,7842% do esforço computacional da BG. Logo, para esse exemplo, evidencia-se uma redução de 85,2158% de operações em relação à técnica tradicional. Pela TABELA 13, nota-se que para a maioria dos conjuntos essa redução foi superior a 70%.

TABELA 13 - COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *MUSHROOM*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	312	100	28,6501	-	71,3499
A2	1089	161	100	14,7842	-	85,2158
A3	1089	292	100	26,8136	-	73,1864
A4	1089	253	100	23,2323	-	76,7677
A5	1089	258	100	23,6915	-	76,3085
A6	1089	142	100	13,0395	-	86,9605
A7	1089	257	100	23,5996	-	76,4004
A8	1089	263	100	24,1506	-	75,8494
A9	1089	140	100	12,8558	-	87,1442
A10	1089	266	100	24,4261	-	75,5739
A11	1089	290	100	26,6299	-	73,3701
A12	1089	335	100	30,7622	-	69,2378
A13	1089	247	100	22,6814	-	77,3186
A14	1089	270	100	24,7934	-	75,2066
A15	1089	147	100	13,4986	-	86,5014
A16	1089	335	100	30,7622	-	69,2378
A17	1089	99	100	9,0909	-	90,9091
A18	1089	346	100	31,7723	-	68,2277
A19	1089	141	100	12,9477	-	87,0523
A20	1089	261	100	23,9669	-	76,0331
A21	1089	296	100	27,1809	-	72,8191
A22	1089	222	100	20,3857	-	79,6143
A23	1089	331	100	30,3949	-	69,6051
A24	1089	99	100	9,0909	-	90,9091
A25	1089	142	100	13,0395	-	86,9605
A26	1089	100	100	9,1827	-	90,8173
A27	1089	104	100	9,5500	-	90,4500
A28	1089	301	100	27,6400	-	72,3600
A29	1089	244	100	22,4059	-	77,5941
A30	1089	268	100	24,6097	-	75,3903

FONTE: A autora (2016).

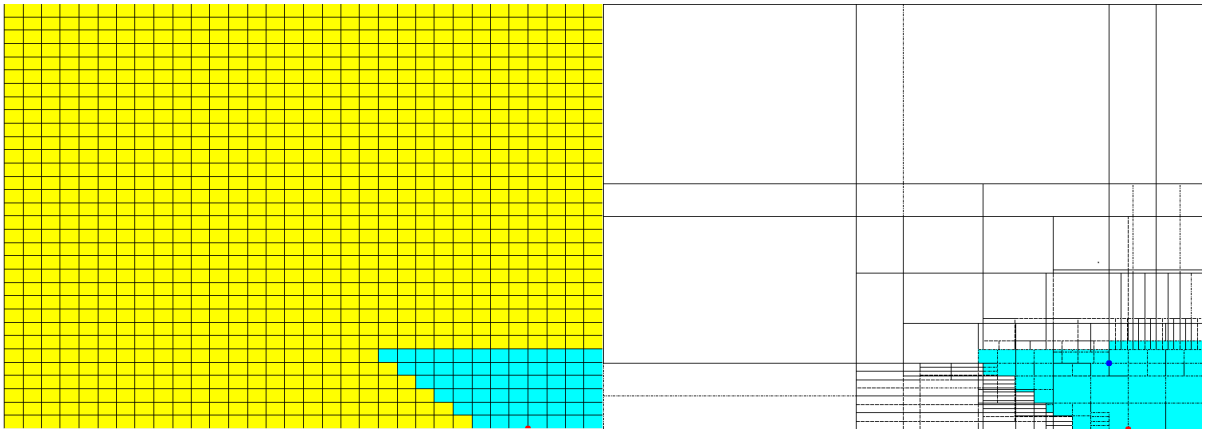
Apesar deste trabalho não apontar em seus resultados o tempo computacional despendido por ambos os métodos, enfatiza-se que essa medida é diretamente proporcional à quantidade de cálculos efetuados. Desta forma, destaca-se que enquanto a BG demorou cerca de 9 horas e 50 minutos para indicar a solução de cada conjunto de dados, o GQ encontrou a resposta em torno de 50 minutos. Em

alguns casos, o tempo computacional da BG superou a marca de 10 horas e 15 minutos, ao passo que o GQ levou no máximo 1 hora para resolver o problema.

No entanto, é importante ressaltar que o tempo gasto para calcular cada combinação de parâmetro (C, γ) não é igual entre si. O cálculo da validação cruzada *5-fold* (QUADRO 15) para alguns parâmetros localizados nas regiões de *underfitting* e *overfitting* é bem mais demorado do que o dos pertencentes à boa região. Logo, por desconsiderar muitos dos pontos externos a essa área, o método *grid-quadtree* torna-se efetivamente mais rápido que a busca por *grid*. Para ilustrar essa explicação, a FIGURA 44 apresenta as soluções gráficas fornecidas pela BG (à esquerda) e GQ (à direita) para o conjunto A16 da base de dados *Mushroom.*

Na FIGURA 44, as interseções dos quadrados (à esquerda) e dos quadrantes (à direita) representam respectivamente os 1089 e 335 pares de parâmetros avaliados pela BG e pelo GQ. Em ambas as soluções, a área em azul claro demonstra a boa região e os pontos destacados em vermelho consistem na solução ótima (para este exemplo, iguais). Exclusivamente na parte direita da FIGURA 44, o ponto azul escuro caracteriza a solução inicial da *quadtree.*

FIGURA 44 - SOLUÇÃO GRÁFICA DO CONJUNTO *MUSHROOM* A16 FORNECIDA PELA BG E GQ



FONTE: A autora (2016).

Com o intuito de avaliar a qualidade das respostas sugeridas pela BG e pelo GQ, parâmetros da TABELA 12, esses foram empregados na classificação dos conjuntos de teste. Assim, a TABELA 14 mostra a quantidade de acertos (QA) e a acurácia atingida pelo SVC para a base *Mushroom*, quando usados tais parâmetros.

TABELA 14 - QA E ACURÁCIA DO SVC PARA O CONJUNTO *MUSHROOM*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1 ... A30	1625/1625	100	1625/1625	100

FONTE: A autora (2016).

Pela TABELA 14, verifica-se que para todos os conjuntos estudados a acurácia do SVC foi igual a 100%, tanto com as sugestões da busca por *grid* quanto com as do *grid-quadtree.* Desta forma, evidencia-se que o SVC ajustado com os parâmetros do GQ (SVC-GQ) é capaz de fornecer soluções tão boas quanto o SVC treinado com os parâmetros da BG (SVC-BG). No QUADRO 25, mostra-se a matriz de confusão referente aos resultados da TABELA 14.

QUADRO 25 - MATRIZ DE CONFUSÃO DA BG E GQ PARA OS CONJUNTOS *MUSHROOM* A1 A A30

	CLASSE 1	CLASSE 2	TOTAL
CLASSE 1	783 (100%)	0 (0%)	783
CLASSE 2	0 (0%)	842 (100%)	842

FONTE: A autora (2016).

Diferentemente do que foi observado para a base de dados *Muhsroom,* em termos de quantidade de acertos e acurácia do SVC, os resultados pertinentes à base *Liver Disorders* foram, dentre as do grupo 1, os menos expressivos. Contudo, enfatiza-se que essa menor eficiência se deu tanto para o SVC-BG quanto para o SVC-GQ. A possível explicação que se faz acerca da inferioridade desses desempenhos é que o kernel gaussiano, neste caso, não é a melhor opção para representar o comportamento dos dados. Diante disso, sugere-se a avaliação de outro tipo de kernel para essa aplicação.

Examinando-se a TABELA 15, que mostra as características dos parâmetros encontrados pela BG e pelo GQ para a base *Liver Disorders,* a inadequação do kernel fica ainda mais evidente quando reparados os percentuais de vetores suporte relacionados aos pares (C, γ). Pela TABELA 15, percebe-se que além das quantidades de VS serem relativamente altas, em média superiores a 50%, grande parte dos VS consiste em erros, isto é, VS*Bound*.

TABELA 15 - PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *LIVER-DISORDERS*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG	11,3137	0,2500	72,4638	168	60,8696	152
	GQ	17,2602	0,2013	72,8261	165	59,7826	150
A2	BG	256,0000	0,5000	73,9130	126	45,6522	77
	GQ	139,5850	0,7071	73,9130	128	46,3768	75
A3	BG	90,5097	0,0625	74,6377	159	57,6087	147
	GQ	35,6604	0,8052	74,6377	141	51,0870	94
A4	BG	128,0000	0,7071	72,1014	137	49,6377	81
	GQ	197,4030	0,5000	72,4638	137	49,6377	88
A5	BG	22,6274	1,0000	75,7246	143	51,8116	97
	GQ	14,6721	0,4204	75,0000	153	55,4348	130
A6	BG	90,5097	0,1768	75,7246	149	53,9855	129
	GQ	35,6604	0,2500	76,0870	152	55,0725	132
A7	BG	128,0000	0,0625	73,1884	157	56,8841	142
	GQ	189,0338	0,0526	72,8261	156	56,5217	140
A8	BG	128,0000	0,1768	74,6377	147	53,2609	126
	GQ	90,5097	0,2500	73,9130	145	52,5362	119
A9	BG	256,0000	0,1250	74,2754	147	53,2609	124
	GQ	45,2548	0,5000	73,5507	144	52,1739	108
A10	BG	64,0000	0,3536	77,1739	132	47,8261	102
	GQ	92,4916	0,3071	77,1739	129	46,7391	101
A11	BG	256,0000	0,0625	76,4493	143	51,8116	127
	GQ	66,1136	0,3497	76,0870	135	48,9130	106
A12	BG	181,0193	1,0000	72,4638	127	46,0145	55
	GQ	10,0431	2,7678	72,4638	147	53,2609	67
A13	BG	181,0193	0,0625	70,6522	156	56,5217	141
	GQ	50,4314	0,1250	71,0145	159	57,6087	144
A14	BG	256,0000	0,0313	73,1884	157	56,8841	145
	GQ	4,0000	5,6569	73,1884	170	61,5942	65
A15	BG	8,0000	1,4142	73,9130	152	55,0725	103
	GQ	197,4030	0,1487	73,5507	144	52,1739	119
A16	BG	256,0000	0,0313	73,1884	160	57,9710	147
	GQ	173,3447	0,0442	72,8261	159	57,6087	146
A17	BG	256,0000	0,1768	72,4638	147	53,2609	121
	GQ	117,3765	0,2500	72,1014	150	54,3478	122
A18	BG	22,6274	0,5000	72,4638	154	55,7971	122
	GQ	53,8174	0,3536	72,1014	149	53,9855	117
A19	BG	64,0000	0,7071	76,4493	137	49,6377	88
	GQ	47,2584	0,8052	76,8116	138	50,0000	89
A20	BG	90,5097	0,3536	75,0000	141	51,0870	106
	GQ	96,5865	0,3536	75,3623	141	51,0870	105

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A21	BG	16,0000	0,5000	75,7246	148	53,6232	120
	GQ	165,9955	0,5000	74,6377	129	46,7391	84
A22	BG	128,0000	0,1768	75,0000	144	52,1739	120
	GQ	56,2001	0,2500	75,0000	147	53,2609	124
A23	BG	5,6569	0,2500	71,3768	175	63,4058	160
	GQ	139,5850	0,3536	69,9275	146	52,8986	107
A24	BG	64,0000	0,2500	72,1014	150	54,3478	124
	GQ	92,4916	0,2243	72,1014	148	53,6232	122
A25	BG	45,2548	0,3536	74,2754	147	53,2609	119
	GQ	139,5850	0,2394	74,2754	141	51,0870	112
A26	BG	256,0000	0,0313	72,8261	161	58,3333	151
	GQ	50,4314	0,0884	72,8261	164	59,4203	151
A27	BG	181,0193	0,0442	74,2754	163	59,0580	152
	GQ	36,4412	0,4391	73,5507	148	53,6232	117
A28	BG	90,5097	0,1250	71,7391	152	55,0725	134
	GQ	36,4412	0,2500	71,0145	153	55,4348	129
A29	BG/GQ	256,0000	0,1768	74,2754	145	52,5362	117
A30	BG	45,2548	0,1768	72,4638	151	54,7101	134
	GQ	36,4412	0,1487	72,8261	156	56,5217	141

FONTE: A autora (2016)

Para a base a *Liver Disorders,* somente em uma situação os métodos BG e GQ encontraram a mesma solução, que foi no conjunto A29 (TABELA 15). Nas demais, determinaram-se diferentes pares (C, γ), mas que possuem características (taxa VC e número de VS) relativamente semelhantes.

Todavia, embora os parâmetros fornecidos pela BG e GQ não tenham resultado em altas acurácias do SVC (TABELA 17), o *grid-quadtree* novamente efetuou menos operações para determinar os pares (C, γ). Na TABELA 16, comparam-se as quantidades de operações realizadas por cada um dos métodos, para a base *Liver Disorders*, da onde nota-se que a menor das reduções de operações proporcionada pelo GQ foi de 74,1965%, que é uma economia bastante expressiva.

TABELA 16 - COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *LIVER-DISORDERS*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	230	100	21,1203	-	78,8797
A2	1089	169	100	15,5188	-	84,4812
A3	1089	195	100	17,9063	-	82,0937
A4	1089	200	100	18,3655	-	81,6345
A5	1089	171	100	15,7025	-	84,2975
A6	1089	213	100	19,5592	-	80,4408
A7	1089	210	100	19,2838	-	80,7163
A8	1089	152	100	13,9578	-	86,0422
A9	1089	132	100	12,1212	-	87,8788
A10	1089	281	100	25,8035	-	74,1965
A11	1089	162	100	14,8760	-	85,1240
A12	1089	207	100	19,0083	-	80,9917
A13	1089	132	100	12,1212	-	87,8788
A14	1089	133	100	12,2130	-	87,7870
A15	1089	218	100	20,0184	-	79,9816
A16	1089	162	100	14,8760	-	85,1240
A17	1089	186	100	17,0799	-	82,9201
A18	1089	153	100	14,0496	-	85,9504
A19	1089	166	100	15,2433	-	84,7567
A20	1089	132	100	12,1212	-	87,8788
A21	1089	161	100	14,7842	-	85,2158
A22	1089	170	100	15,6107	-	84,3894
A23	1089	142	100	13,0395	-	86,9605
A24	1089	251	100	23,0487	-	76,9513
A25	1089	228	100	20,9366	-	79,0634
A26	1089	131	100	12,0294	-	87,9706
A27	1089	155	100	14,2332	-	85,7668
A28	1089	134	100	12,3049	-	87,6951
A29	1089	231	100	21,2121	-	78,7879
A30	1089	240	100	22,0386	-	77,9614

FONTE: A autora (2016).

A TABELA 17 apresenta as quantidades de acertos e as acurácias atingidas pelo SVC quando usados os parâmetros da TABELA 15. Conforme mencionado, tanto o desempenho do SVC-BG quanto do SVC-GQ para a base *Liver Disorders* foram de regulares a não satisfatórios. Neste trabalho, considera-se baixa performance as acurácias inferiores a 70%, como por exemplo a do conjunto A10 (60,8696%).

TABELA 17 - QA E ACURÁCIA DO SVC PARA O CONJUNTO *LIVER-DISORDERS*

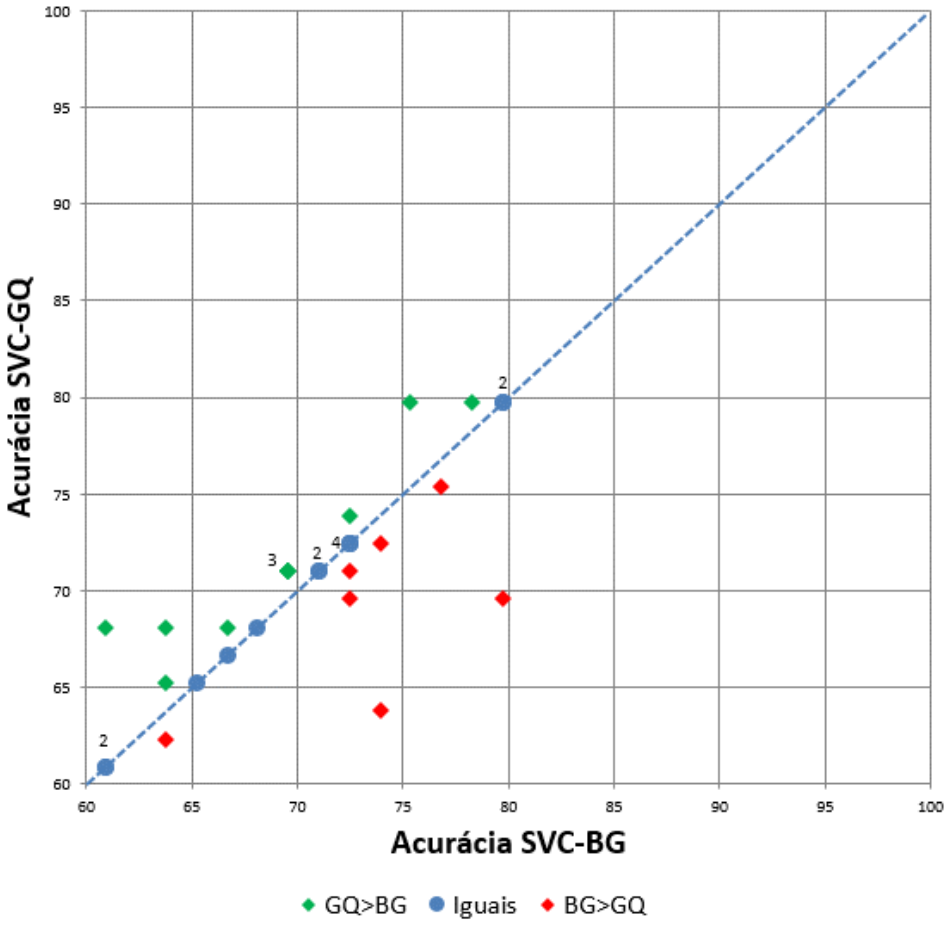
CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	53/69	76,8116	52/69	75,3623
A2	47/69	68,1159	47/69	68,1159
A3	48/69	69,5652	49/69	71,0145
A4	48/69	69,5652	49/69	71,0145
A5	42/69	60,8696	47/69	68,1159
A6	46/69	66,6667	47/69	68,1159
A7	50/69	72,4638	50/69	72,4638
A8	50/69	72,4638	48/69	69,5652
A9	50/69	72,4638	49/69	71,0145
A10	42/69	60,8696	42/69	60,8696
A11	44/69	63,7681	43/69	62,3188
A12	42/69	60,8696	42/69	60,8696
A13	55/69	79,7101	55/69	79,7101
A14	51/69	73,9130	44/69	63,7681
A15	44/69	63,7681	47/69	68,1159
A16	50/69	72,4638	50/69	72,4638
A17	55/69	79,7101	55/69	79,7101
A18	54/69	78,2609	55/69	79,7101
A19	45/69	65,2174	45/69	65,2174
A20	49/69	71,0145	49/69	71,0145
A21	44/69	63,7681	45/69	65,2174
A22	49/69	71,0145	49/69	71,0145
A23	52/69	75,3623	55/69	79,7101
A24	51/69	73,9130	50/69	72,4638
A25	46/69	66,6667	46/69	66,6667
A26	50/69	72,4638	50/69	72,4638
A27	55/69	79,7101	48/69	69,5652
A28	50/69	72,4638	51/69	73,9130
A29	50/69	72,4638	50/69	72,4638
A30	48/69	69,5652	49/69	71,0145

FONTE: A autora (2016).

Pela TABELA 17, verifica-se que em 13 dos conjuntos avaliados, o SVC-BG e o SVC-GQ tiveram o mesmo desempenho. Já em 10 dos conjuntos o SVC-GQ foi superior ao SVC-BG e nos 7 restantes o contrário. O GRÁFICO 1 ilustra essas relações de igualdade e superioridade de acurácia, onde a cor verde refere-se as melhores situações do SVC-GQ e a vermelha a do SVC-BG. Ainda no GRÁFICO 1, a linha azul aponta os casos em que ambos apresentaram igual performance. Já a

numeração ao lado dos pontos identifica a quantidade de vezes que determinado resultado se repetiu. Por exemplo, conforme se vê pela TABELA 17 e pelo GRÁFICO 1, em duas ocasiões o SVC-BG e SVC-GQ obtiveram simultaneamente a acurácia 79,7171%, que foram para os conjuntos A13 e A17.

GRÁFICO 1 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *LIVER DISORDERS*



FONTE: A autora (2016).

Os QUADROS 26 e 27 representam respectivamente as matrizes de confusão do conjunto A10 e A17, referentes às piores (60,8696%) e melhores (79,7101%) performances do SVC-GQ e SVC-BG para a base *Liver Disorders*. Por meio dessas matrizes, evidencia-se que em ambas as condições, os dois métodos foram capazes de melhor reconhecer os padrões pertencentes à classe 2.

QUADRO 26 - MATRIZ DE CONFUSÃO DA BG E GQ PARA O CONJUNTO *LIVER DISORDER*S A10

	CLASSE 1	CLASSE 2	TOTAL
CLASSE 1	15 (51,72%)	14 (48,28%)	29
CLASSE 2	13 (32,50%)	27 (67,50%)	40

FONTE: A autora (2016).

QUADRO 27 - MATRIZ DE CONFUSÃO DA BG E GQ PARA O CONJUNTO *LIVER DISORDERS* A17

	CLASSE 1	CLASSE 2	TOTAL
CLASSE 1	22 (75,86%)	7 (24,14%)	29
CLASSE 2	7 (17,50%)	33 (82,50%)	40

FONTE: A autora (2016).

Em função da grande quantidade de resultados obtidos nesta etapa do trabalho, decidiu-se por apresentar as tabelas de respostas e os gráficos de acurácia das demais bases de dados do QUADRO 21 no apêndice desta tese. Contudo, destaca-se que a maneira de interpretá-los é semelhante ao que foi feito para a *Mushroom* e *Liver Disorders.*

No entanto, para facilitar a visualização de todos resultados alcançados, as TABELAS 18 a 21 resumem as principais medidas estatísticas dos conjuntos do grupo 1, no que diz respeito aos indicadores de desempenho: taxa de validação cruzada e percentual de vetores suporte relacionados aos parâmetros fornecidos pela BG e pelo GQ, acurácia do SVC-BG e do SVC-GQ, e redução de operações proporcionadas pelo *grid-quadtree*. Tais medidas foram calculadas considerando-se os 30 conjuntos de cada base.

Pela TABELA 18, que trata da taxa VC, percebe-se que o método GQ determinou, na média, parâmetros (C, γ) com menor taxa de validação cruzada que a BG. Todavia, verifica-se que a maior diferença relativa entre esses valores foi de 0,26% (*Liver Disorders* e *Iris*).

Por outro lado, apesar das menores taxas VC, os pares (C, γ) estipulados pelo GQ possuem, em média, menor percentual de vetores suporte, conforme mostra a TABELA 19. Isso evidencia melhor capacidade de generalização proporcionada pelos parâmetros do GQ, fato que pode ser constatado na TABELA 20 (referente às acurácias).

Assim, na TABELA 20 (linha dos valores médios), observa-se que em cinco das bases estudadas o SVC-GQ apresentou um desempenho superior ao do SVC-BG e em duas delas igual. Para os casos de inferioridade do SVC-GQ, a diferença relativa entre as acurácias médias não foi maior que 0,14% (*Liver Disorders).*

TABELA 18 – RESUMO ESTATÍSTICO DA TAXA VC (%) DOS PARÂMETROS DETERMINADOS PELA BG E PELO GQ

MÉT.	MEDIDA ESTAT.	*MUSHROOM*	*LIVER DISORDERS*	*IONOSPHERE*	*BREAST CANCER*	*DIABETES*	*CIRCLE AND SQUARE*	*IRIS*	*SVMGUIDE2*	*VEHICLE*	*SEGMENT*
BG	Média	100,0000	73,8044	95,5990	97,3077	77,8773	99,4542	97,5833	84,9201	84,9138	97,2547
	Mínimo	100,0000	70,6522	94,3060	96,8864	75,8958	98,8750	95,0000	82,1086	83,3087	96,8074
	Q1	100,0000	72,4638	95,0178	97,0696	77,3616	99,3750	96,6667	84,3450	84,5273	97,1320
	Mediana	100,0000	73,9130	95,7295	97,2527	77,9316	99,5000	97,5000	84,6645	84,7858	97,2403
	Q3	100,0000	74,9094	96,0854	97,4359	78,4609	99,5000	98,3333	85,7828	85,3398	97,3350
	Máximo	100,0000	77,1739	97,1530	97,8022	80,6189	99,7500	99,1667	87,8594	87,2969	97,8355
	Desvio padrão	0,0000	1,6457	0,7208	0,3085	0,9234	0,2063	0,9374	1,4174	0,8619	0,2238
GQ	Média	100,0000	73,6111	95,4448	97,2161	77,7362	99,4125	97,3333	84,7497	84,8400	97,1879
	Mínimo	100,0000	69,9275	93,9502	96,5201	75,8958	98,7500	95,0000	81,7891	83,3087	96,8074
	Q1	100,0000	72,5544	94,7509	96,8864	77,0765	99,2813	96,6667	83,7860	84,3427	97,0779
	Mediana	100,0000	73,5507	95,7295	97,2527	77,8502	99,3750	97,5000	84,6645	84,7858	97,1861
	Q3	100,0000	74,6377	95,7295	97,4359	78,2981	99,5000	97,5000	85,3035	85,3398	97,2944
	Máximo	100,0000	77,1739	96,7972	97,8022	80,9446	99,8750	99,1667	87,8594	87,2969	97,8355
	Desvio padrão	0,0000	1,7129	0,7897	0,3738	1,0222	0,2278	0,9639	1,4063	0,8600	0,2176

FONTE: A autora (2016).

TABELA 19 – RESUMO ESTATÍSTICO DO NÚMERO DE VS (%) DOS PARÂMETROS DETERMINADOS PELA BG E PELO GQ

MÉT.	MEDIDA ESTAT.	*MUSHROOM*	*LIVER DISORDERS*	*IONOSPHERE*	*BREAST CANCER*	*DIABETES*	*CIRCLE AND SQUARE*	*IRIS*	*SVMGUIDE2*	*VEHICLE*	*SEGMENT*
BG	Média	2,6932	54,0459	30,9609	9,0721	45,1086	5,4125	19,6667	45,0053	37,9665	16,9751
	Mínimo	2,4311	45,6522	16,7260	5,8608	39,5765	2,8750	10,8333	33,2268	35,4505	12,9329
	Q1	2,6043	51,9022	21,3523	7,1887	42,3046	3,6250	11,6667	40,2556	36,3737	14,8268
	Mediana	2,6850	53,8044	25,6228	8,2418	44,3811	4,4375	17,0834	45,5272	37,4447	16,6667
	Q3	2,7812	56,7935	41,9929	10,5311	47,7199	6,8438	23,1250	49,0416	39,0695	19,0206
	Máximo	3,0005	63,4058	46,2633	17,0330	53,7459	11,3750	45,0000	57,5080	43,4269	21,1039
	Desvio padrão	0,1325	4,0403	10,7027	2,7041	3,6132	2,4074	9,0671	6,5603	1,8766	2,4784
GQ	Média	2,6994	53,3696	28,9917	8,0647	43,2193	4,7958	18,6667	44,6539	37,9518	15,7846
	Mínimo	2,4158	46,3768	16,7260	4,9451	38,9251	2,8750	10,8333	32,9074	35,8937	13,4740
	Q1	2,6043	51,0870	19,6620	6,5019	41,8567	3,6563	11,6667	38,4186	36,6322	14,5428
	Mediana	2,7081	53,2609	24,9111	7,5092	42,7525	4,4375	14,5834	42,0128	37,6662	15,6115
	Q3	2,7850	55,4348	39,4129	8,2418	44,0961	5,1250	22,2917	49,6805	38,8479	16,9102
	Máximo	3,0005	61,5942	46,6192	17,9487	49,6743	9,5000	47,5000	61,3419	42,5406	19,7511
	Desvio padrão	0,1366	3,7883	10,5592	2,4290	2,2500	1,5304	9,4777	8,1275	1,6832	1,5712

FONTE: A autora (2016).

TABELA 20 – RESUMO ESTATÍSTICO DA ACURÁCIA (%) OBTIDA PELOS MÉTODOS SVC-BG E SVC-GQ

MÉT.	MEDIDA ESTAT.	*MUSHROOM*	*LIVER DISORDERS*	*IONOSPHERE*	*BREAST CANCER*	*DIABETES*	*CIRCLE AND SQUARE*	*IRIS*	*SVMGUIDE2*	*VEHICLE*	*SEGMENT*
SVC-BG	Média	100,0000	70,5314	94,4762	96,8370	76,0390	98,8333	96,0000	83,2051	84,6746	97,1717
	Mínimo	100,0000	60,8696	88,5714	94,1606	68,1818	97,0000	90,0000	74,3590	79,8817	95,0216
	Q1	100,0000	66,6667	92,8571	95,8029	74,1883	98,5000	94,1667	79,4872	83,4320	96,9697
	Mediana	100,0000	71,7392	94,2857	97,0803	76,6234	99,0000	96,6667	83,3333	84,6154	97,1861
	Q3	100,0000	73,5507	95,7143	97,8102	78,4091	99,3750	96,6667	86,8590	86,2426	97,6190
	Máximo	100,0000	79,7101	98,5714	100,0000	82,4675	100,0000	100,0000	91,0256	88,7574	98,2684
	Desvio padrão	0,0000	5,5979	2,3349	1,4514	3,4248	0,6865	2,5371	4,2367	2,1940	0,6529
SVC-GQ	Média	100,0000	70,4348	94,8095	96,8370	76,5368	98,8500	96,2222	83,5043	84,6351	97,1068
	Mínimo	100,0000	60,8696	90,0000	94,1606	67,5325	97,0000	90,0000	75,6410	80,4734	95,0216
	Q1	100,0000	68,1159	92,8571	95,6204	74,8377	98,5000	96,6667	80,1282	83,4320	96,7532
	Mediana	100,0000	71,0145	94,2857	97,0803	76,9481	99,0000	96,6667	83,3333	84,3196	97,1861
	Q3	100,0000	72,4638	96,7858	98,3576	78,5714	99,3750	96,6667	85,8974	86,3905	97,4026
	Máximo	100,0000	79,7101	98,5714	99,2701	83,1169	100,0000	100,0000	91,0256	88,1657	98,0519
	Desvio padrão	0,0000	5,2129	2,2969	1,5375	3,4264	0,6318	2,5869	4,1745	2,0991	0,6002

FONTE: A autora (2016).

TABELA 21 – RESUMO ESTATÍSTICO DA REDUÇÃO DE OPERAÇÕES (%) PROPORCIONADA PELO MÉTODO *GRID-QUADREE*

MEDIDA ESTATÍSTICA	*MUSHROOM*	*LIVER DISORDERS*	*IONOSPHERE*	*BREAST CANCER*	*DIABETES*	*CIRCLE AND SQUARE*	*IRIS*	*SVMGUIDE2*	*VEHICLE*	*SEGMENT*
Média	78,8124	83,3272	85,1240	83,7833	83,8415	80,0245	79,4399	84,5271	80,9060	84,1720
Mínimo	68,2277	74,1965	78,6961	47,1993	55,7392	66,2994	59,3205	70,0643	71,5335	66,6667
Q1	73,2323	80,5097	82,2544	82,3692	83,1267	77,9844	75,2985	82,1626	79,9357	82,8972
Mediana	76,3545	84,4353	87,0524	85,2158	86,2719	79,2011	81,4968	85,675	81,3131	85,0322
Q3	86,8457	86,0193	87,764	87,7411	87,7870	84,2516	85,2847	87,5804	82,6447	87,2819
Máximo	90,9091	87,9706	88,1543	91,0009	88,2461	87,787	88,4298	91,0927	87,9706	88,0624
Desvio padrão	7,4032	3,8024	3,4727	7,6084	6,55	5,0727	8,0708	4,4808	3,5310	4,4308

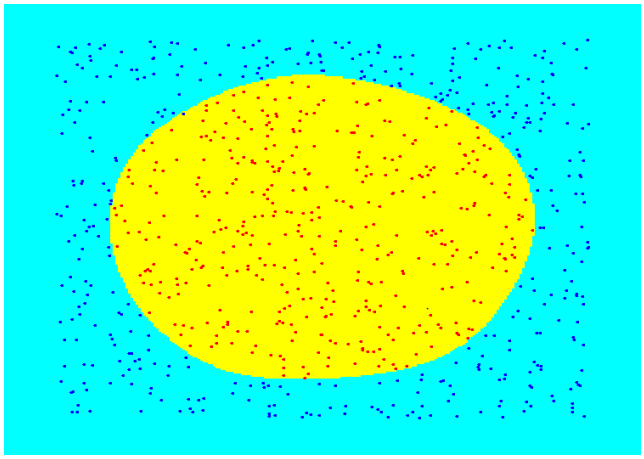
FONTE: A autora (2016).

Já pela TABELA 21, que destaca as medidas estatísticas referentes a redução de operações do método *grid-quadtree*, verifica-se que em média o GQ foi capaz de encontrar parâmetros reduzindo pelo menos 78,8124% (*Mushroom*) de operações da busca por *grid.* Se observados isoladamente os valores mínimos e máximos da TABELA 21, dentre os 300 conjuntos avaliados, tem-se que a menor redução de operações proporcionada pelo GQ foi de 47,1993% (*Breast Cancer*) e a maior de 91,0927% (*Svmguide 2).* Logo, constata-se que o pior dos resultados do GQ, em termos de redução de operações, já é bastante expressivo. Ainda, conforme se vê pela TABELA 20, essas reduções foram realizadas sem, no entanto, afetar a acurácia do SVC-GQ.

Para finalizar as explicações pertinentes ao grupo 1, ilustra-se separação das classes -1 e +1 da base *Circle and Square*, correspondente ao treinamento do conjunto A9, cujo hiperplano foi determinado com os parâmetros (256,0000; 2,0000), sugeridos tanto pela BG quanto pelo GQ (TABELA 34 do apêndice). Escolheu-se o A9 em especial, pois foi nesse conjunto em que o SVC-GQ e SVC-BG obtiveram a menor acurácia na classificação dos dados de teste (97,0000%), segundo mostra a TABELA 36 do apêndice. Dessa forma, desejou-se mostrar as características da função que menos padrões acertou (194/200).

Pela FIGURA 45, nota-se que o hiperplano separador do conjunto A9 identifica sem erros as classes positiva (pontos vermelhos) e negativa (pontos azuis), discriminado adequadamente o círculo dentro do quadrado, conforme indica o seu nome em inglês (*Circle and Square*). Essa base, por possuir apenas duas características, é a única do grupo 1 que pode ter seu hiperplano ilustrado no plano.

FIGURA 45 - SEPARAÇÃO DAS CLASSES DO CONJUNTO A9 DA BASE *CIRCLE AND SQUARE*



FONTE: A autora (2016).

5.2 RESULTADOS COMPUTACIONAIS DO GRUPO 2

Os resultados do grupo 2 referem-se às bases de dados descritas no QUADRO 22, cujos conjuntos de treinamento e de teste já se encontravam devidamente separados no LIBSVM. Na TABELA 22, descrevem-se as características dos parâmetros encontrados pela busca por *grid* e pelo *grid-quadtree* para essas bases, onde as cinco primeiras são binárias e as cinco últimas multiclasse. Portanto, da A1A a W1A constam resultados de VS*Bound* e para as demais não.

TABELA 22 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA OS CONJUNTOS DO GRUPO 2

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1A	BG	2,8284	0,0442	83,2399	549	34,2056	392
	GQ	2,5657	0,0414	83,2399	546	34,0187	410
Splice	BG	5,6569	0,0156	87,9000	475	47,5000	44
	GQ	6,7272	0,0143	87,9000	469	46,9000	40
Svmguide 1	BG	11,3137	8,0000	96,9569	288	9,3234	136
	GQ	7,1788	1,4142	96,9246	257	8,3198	228
Svmguide 3	BG/GQ	128,0000	0,0884	84,3926	390	31,3757	231
W1A	BG	256,0000	0,0078	98,1428	231	9,3258	20
	GQ	256,0000	0,0066	98,1833	228	9,2047	20
DNA	BG	2,8284	0,0442	96,0000	1290	64,5000	-
	GQ	1,0000	0,0313	95,6500	1104	55,2000	-
Satimage	BG	5,6569	1,4142	92,4690	1528	34,4532	-
	GQ	3,2210	1,1764	92,4014	1431	32,2661	-
Svmguide 4	BG	256,0000	1,0000	78,3333	158	52,6667	-
	GQ	239,8935	1,0671	78,0000	158	52,6667	-
Pendigits	BG	5,6569	1,0000	99,7198	855	11,4091	-
	GQ	6,8745	0,9170	99,7198	831	11,0889	-
Vowel	BG	2,8284	11,3137	99,4318	330	62,5000	-
	GQ	4,9674	6,8745	99,2424	325	61,5530	-

FONTE: A autora (2016).

Pela TABELA 22, percebe-se que apenas para o conjunto *Svmguide* 3 os métodos BG e GQ determinaram o mesmo par de parâmetros (C, γ). Nos demais, ambos encontraram soluções próximas umas das outras, como por exemplo (256,0000; 0,0078) e (256,0000; 0,0066) para o conjunto W1A, mas relacionadas à diferentes taxas VC e/ou quantidades de VS.

No que diz respeito à taxa de validação cruzada, somente em um dos conjuntos (W1A) o GQ forneceu parâmetros com maior taxa VC que a BG. No restante, as taxas foram menores (em cinco dos conjuntos) ou iguais (em três dos casos). Todavia, nas situações de inferioridade do método proposto, a máxima diferença relativa entre as taxas da BG e do GQ foi de 0,37% (DNA). Em contrapartida, os percentuais de vetores suporte dos parâmetros do GQ foram todos menores, com exceção do conjunto *Svmguide 4* e *Svmguide 3* em que ambos foram iguais.

Em termos de quantidade de operações, novamente o GQ efetuou menos avaliações de parâmetros que a BG para encontrar a solução, sendo em torno de 11,2948 % a 28,2828% do esforço da técnica tradicional. Pela TABELA 23, que compara a quantidade de operações de cada método, nota-se que a maior economia de cálculos realizada pelo GQ foi de 88,7052% (DNA) e a menor foi de 71,7172 (A1A).

TABELA 23 – COMPARAÇÃO DO NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA OS CONJUNTOS DO GRUPO 2

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1A	1089	308	100	28,2828	-	71,7172
Splice	1089	161	100	14,7842	-	85,2158
Svmguide 1	1089	137	100	12,5803	-	87,4197
Svmguide 3	1089	132	100	12,1212	-	87,8788
W1A	1089	186	100	17,0799	-	82,9201
DNA	1089	123	100	11,2948	-	88,7052
Satimage	1089	137	100	12,5803	-	87,4196
Svmguide 4	1089	292	100	26,8136	-	73,1864
Pendigits	1089	158	100	14,5087	-	85,4913
Vowel	1089	232	100	21,3040	-	78,6960

FONTE: A autora (2016).

Dentre as bases do grupo 2, destaca-se a *Pendigits* que, tal qual a *Mushroom,* exige muito tempo de processamento para analisar todo o espaço de busca de parâmetros devido as suas dez classes e grande dimensão (7494 dados de treinamento e 16 características). Desta forma, a redução de operações de 85,4913%, para a *Pendigits,* proporcionada pela GQ impactou diretamente no tempo computacional da tarefa, pois, enquanto a BG demorou cerca de 7 horas e 40 minutos para determinar a solução do problema, o método proposto levou apenas 30 minutos. Ainda, conforme se vê pela TABELA 24, que mostra a quantidade de acertos e

acurácia do SVC, para essa base, o SVC-GQ teve um desempenho ligeiramente superior ao do SVC-BG (98,4848% contra 98,4563%), pois acertou um padrão a mais. Logo, neste caso, o GQ, além de fornecer parâmetros mais rapidamente, o fez com excelente qualidade.

TABELA 24 – QA E ACURÁCIA DO SVC PARA OS CONJUNTOS DA ETAPA 2

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1A	26106/30956	84,3326	26101/30956	84,3164
Splice	1961/2175	90,1609	1957/2175	89,9770
Svmguide 1	3862/4000	96,5500	3876/4000	96,9000
Svmguide 3	34/41	82,9268	34/41	82,9268
W1A	46150/47272	97,6265	46172/47272	97,6730
DNA	1123/1186	94,6880	1127/1186	95,0253
Satimage	1838/2000	91,9000	1845/2000	92,2500
Svmguide 4	247/312	79,1667	247/312	79,1667
Pendigits	3444/3498	98,4563	3445/3498	98,4848
Vowel	275/462	59,5238	290/462	62,7706

FONTE: A autora (2016).

Em relação à acurácia de todo o grupo 2, verifica-se pela TABELA 24 que em seis dos conjuntos avaliados a performance do SVC-GQ foi superior ao do SVC-BG e em dois dos casos igual (*Svmguide 3* e *Svmguide 4*). Para o *Svmguide 3*, a igualdade de desempenhos era esperada uma vez que a BG e o GQ determinaram o mesmo par de parâmetros. Já para as situações de inferioridade do SVC-GQ, observa-se, no entanto, que a diferença entre as quantidades de acertos de ambos os métodos foi de somente 5 a 4 padrões (A1A e *Splice*).

Pela TABELA 24, percebe-se ainda que, com exceção do conjunto *Vowel,* o SVC-GQ apresentou acurácias superiores ou igual à 79,1667% (*Svmguide 4*). Para o *Vowel*, mesmo o SVC-GQ atingindo melhores resultados que o SVC-BG, em ambos os casos fica clara a ocorrência de *overfitting*, ou seja, pequenas taxas de erros no treinamento (inferiores a 1% - TABELA 22) e baixa acurácia no teste (SVC-BG = 59,5238% e SVC-GQ=62,7706%). Contudo, apesar do desempenho ruim, destaca-se que o resultado do GQ está em conformidade com o disponível na literatura.

Dentre os autores que avaliaram a base *Vowel,* estudados nesta tese (QUADRO 30 do apêndice), apenas Hsu e Lin (2002) e Wang, Huang e Cheng (2014)

apresentaram os valores dos parâmetros encontrados por suas técnicas, sendo (16; 0) para o primeiro[17] e (26,86; 0,05) para o segundo. É comum nos trabalhos da área de seleção de modelos do SVC que os autores exponham somente a taxa de validação cruzada evidenciada no treinamento (para a determinação dos parâmetros) e o tempo de execução dessa tarefa. Desta forma, raro são os trabalhos que, além de exibir as suas respostas de parâmetros, verificam a sua qualidade por meio da predição dos dados de teste.

Todavia, apesar dos conjuntos de treinamento e de teste do *Vowel* estarem disponibilizados separadamente no LIBSVM, Wang, Huang e Cheng (2014) usaram outra composição (divisão)[18] dos conjuntos em sua pesquisa. Portanto, para certificar os resultados encontrados pelo GQ, referente à base *Vowel,* avaliou-se somente o comportamento do par (C, γ) de Hsu e Lin (2002). Para tal, empregaram-se seus parâmetros na classificação do conjunto de teste da *Vowel* (QUADRO 22) e compararam-se seus resultados com o do GQ. Nessa verificação, obteve-se para Hsu e Lin (2002) uma acurácia do SVC de 59,7403% (QA= 276/462), que é inferior ao valor do GQ, destacado na TABELA 24.

Nesse contexto, ressalta-se também que, embora não se conheçam os pares (C, γ) e a acurácia do SVC da base *Vowel* para a maioria dos autores do QUADRO 30 do apêndice, Hsu e Lin (2002), Lin *et al.* (2008a), Lin *et al.* (2008b) afirmam que seus parâmetros foram encontrados, nesse caso, com taxas de validação cruzada superiores a 99%. Logo, a taxa VC do GQ (99,2424%) é condizente com tais trabalhos e com o resultado da busca por *grid* (99,4318%).

Para finalizar, explica-se ainda que, ao deparar-se com a situação de *overfitting* da base *Vowel,* analisou-se tanto para o GQ quanto para a BG parâmetros em um espaço de busca com extremos superiores e inferiores a 2^{8} e 2^{-8} (limites do eixo). A finalidade disso foi avaliar se a boa região de parâmetros poderia estar contida em outra área além da já investigada. Entretanto, em virtude de não se obter qualquer melhoria de resultados, entende-se que o kernel gaussiano pode não ser o mais indicado para tal aplicação.

[17] Hsu e Lin (2002) apresentaram para o conjunto *Vowel* pares de parâmetros encontrados por diferentes metodologias de SVC multiclasse. Em consonância com a tese, utilizou-se como referência aquele determinado pelo método “Um contra um”.

[18] Utilizaram 330 dados para treinamento e 660 para teste.

6 CONSIDERAÇÕES FINAIS

A performance do algoritmo *Support Vector Classification* (SVC) depende da correta seleção de seus parâmetros: constante de regularização C, função kernel e suas respectivas variáveis. Dentre os métodos existentes de determinação de parâmetros do SVC, destaca-se a busca por *grid* (BG), que tem por finalidade encontrar a constante C e o parâmetro γ do kernel gaussiano em uma malha (*grid*). A BG, devido a sua simplicidade, é uma das metodologias mais utilizadas entre os usuários do SVC, mas possui elevado custo computacional, pois avalia todas as combinações de parâmetros (C, γ) em seu espaço de busca.

Neste contexto, o objetivo geral desta tese foi propor um método de seleção de parâmetros do SVC, usando o kernel gaussiano, que combinasse a técnica *quadtree* à busca por *grid*, para reduzir o número de operações (cálculos) efetuados pelo *grid* e diminuir seu custo computacional. Para o funcionamento do método proposto, intitulado *grid-quadtree*, utilizou-se uma *quadtree* balanceada e refinada, cujo mecanismo de refinamento foi criado neste trabalho.

A partir do estudo da teoria do problema de seleção de modelos do SVC e das características da *quadtree*, identificou-se a possibilidade de empregar a técnica *quadtree* para desenhar a boa região de parâmetros, evidenciada por Keerthi e Lin (2003), e localizar nessa região um par (C, γ) adequado para o SVC. A ideia fundamental do método GQ foi reduzir o número de operações da BG evitando-se a avaliações desnecessárias de parâmetros situados nas áreas de *underfitting* e *overfitting.*

O método aqui desenvolvido e a busca por *grid* foram implementados utilizando-se a linguagem de programação VB.net e os softwares da biblioteca LIBSVM. Ambos os métodos foram testados usando-se vinte bases de dados referência da área de classificação, as quais foram divididas em dois grupos: um contendo conjuntos de treinamento e de teste separados aleatoriamente e outro com conjuntos previamente definidos.

Umas das grandes vantagens do GQ é que ele permite a seleção de parâmetros do SVC observando-se duas medidas de desempenho: taxa de validação cruzada (VC) e número de vetores suporte (VS). Embora a quantidade de VS seja um indicador muito importante ao SVC, pois reflete a capacidade de generalização do

algoritmo, raros são os métodos de determinação de parâmetros disponíveis na literatura que avaliam tal medida. A busca por *grid*, em sua concepção original, é um exemplo de metodologia que considera apenas a taxa de validação cruzada na sua procura. No entanto, adaptou-se nesta tese o cálculo de VS à BG para que a comparação entre os métodos fosse a mais adequada possível.

Assim, confrontaram-se os resultados do *grid-quadtree* com os da tradicional BG e, a partir dessa análise, avaliou-se o desempenho do método proposto. Para ambos grupos de dados estudados, o GQ encontrou parâmetros com altas de validação cruzada e baixa quantidade de vetores suporte. As taxas médias de VC do GQ para o grupo 1 foram superiores a 73,6111% e a do grupo 2 a 78,0000%, chegando a atingir 100% e 99,7198%, respectivamente.

Apesar dos valores de taxa VC serem normalmente inferiores aos determinados pela BG, a maior diferença relativa entre as taxas de ambos os métodos foi de 0,37%. Contudo, verificou-se que essa inferioridade nas taxas do GQ não afetou a quantidade de acertos do SVC, pois em 11 das 20 bases de dados analisadas a acurácia do SVC-GQ foi superior ao do SVC-BG e em quatro delas iguais. Isso mostra que, mesmo selecionando pares de parâmetros (C, γ) com taxa VC ligeiramente mais baixa, as respostas do GQ propiciam bom desempenho ao SVC.

Ainda, destaca-se que os resultados de taxa VC do GQ, encontrados nesta tese, estão de acordo com os dos vários trabalhos estudados na revisão de literatura. Devido à quantidade de referências pesquisadas e também por não ser objetivo deste trabalho, não se calculou a diferença relativa entre as taxas VC do GQ e de outros métodos além da BG. Todavia, verificam-se que as bases de dados que apresentaram, por exemplo, taxas em torno de 70%, 80% e 90% tiveram o mesmo comportamento para os demais métodos. Isso significa que se o GQ determinou uma taxa VC de aproximadamente 77% para certa base, os demais métodos também encontraram valores próximos a isso, não se observando discrepâncias entre eles.

Em relação ao número de vetores suporte, na maioria dos casos estudados as quantidades de VS vinculadas aos parâmetros do GQ foram menores que 50%. Além disso, em 19 das 20 bases analisadas, o GQ estabeleceu parâmetros com menor percentual de VS que a BG, mostrando que suas repostas fornecem boa capacidade de generalização ao SVC. O único caso em que a generalização não ocorreu, devido a um problema de *overfitting*, foi no conjunto *Vowel.* Entretanto, como

tanto o SVC-BG quanto o SVC-GQ apresentaram baixo desempenho, conclui-se a necessidade de se avaliar, nessa situação, a aplicação de outra função kernel.

A grande contribuição do método GQ foi, sem dúvida, a redução de operações proporcionada pela aplicação da *quadtree*. Em muitos dos casos, verificou-se que o GQ foi capaz de encontrar os mesmos pares de parâmetros (C, γ) que a BG ou melhores realizando muito menos operações. Em relação à técnica tradicional, constatou-se no grupo 1 uma redução média de operações de pelo menos 78,8124%, e no grupo 2 de 71,7172% a 88,7052%. Essa diminuição na quantidade de cálculos efetuados pela *quadtree* acarretou em uma economia de processamento, que para os grandes conjuntos de dados foi superior a sete horas.

Além disso, pela *quadtree* ser uma técnica razoavelmente simples de compreender e não necessitar de parâmetros adicionais, a aplicação do método *grid-quadtree* torna-se acessível aos usuários do SVC, que contam com uma nova opção de ferramenta para selecionar parâmetros.

Por fim, pelas bases de dados avaliadas serem de referência na área de classificação, os resultados do *grid-quadtree* tornam-se ainda mais expressivos. Outro diferencial deste trabalho, diferentemente dos muitos observados na revisão de literatura, é que, além de mostradas as taxas de validação cruzada relacionadas aos parâmetros (C, γ) selecionados, avaliou-se a qualidade dos mesmos por meio da classificação dos conjuntos de teste. A partir dessa análise, conclui-se que usar a *quadtree* como forma de otimizar o tempo computacional do *grid* é viável, já que o GQ encontra parâmetros que conferem alta acurácia ao SVC, efetuando uma menor quantidade de operações.

6.1 SUGESTÕES PARA TRABALHOS FUTUROS

Sugere-se como objeto de pesquisa para trabalhos futuros o estudo da seleção de características dos vetores de treinamento, em conjunto com o método *grid-quadtree*. Conforme mostrado na revisão de literatura, tal tarefa pode ser realizada antes ou simultaneamente à determinação de parâmetros. Logo, recomenda-se a aplicação de uma ferramenta que trabalhe anteriormente ou juntamente ao funcionamento da *quadtree*.

No caso de concomitância, da seleção de características e do par de parâmetros (C, γ), sugere-se a análise de viabilidade do uso da *octree.* A *octree* é uma estrutura, derivada da *quadtree,* destinada à representação de dados tridimensionais. Essa técnica baseia-se na divisão sucessiva do espaço em octantes do mesmo tamanho que, assim como na *quadtree,* realiza a partição do espaço até que todos os octantes estejam inteiramente contidos ou fora da região de interesse. Desta forma, nessa aplicação, recomenda-se que a terceira coordenada do espaço esteja relacionada à determinação de características do vetor.

Ainda nesse tema *quadtree/octree*, sugere-se a extensão do modelo *grid-quadtree* aos problemas de regressão. Pelos mesmos motivos apresentados para o SVC, o kernel gaussiano também é o mais empregado no *Support Vector Regression* (SVR). Dessa maneira, três são os parâmetros de importância para o bom desempenho do SVR: constante de regularização C, γ do kernel gaussiano e margem ε. Portanto, recomenda-se o desenvolvimento de um modelo *grid-octree* para a seleção de parâmetros (C, γ, ε) para o SVR.

Novamente para o SVC, outra sugestão de pesquisa é, após identificar a boa região de parâmetros por meio da técnica *quadtree,* aprimorar o mecanismo de busca pelo parâmetro ótimo (C, γ) situado nessa área. Assim, recomenda-se a elaboração de ferramentas que vasculhem pequenas regiões na proximidade do melhor vértice, determinado pelo método *grid-quadtree*, sem, no entanto, comprometer o número de operações realizadas.

Além das possibilidades anteriores, sugere-se o estudo do ponto de referência (1,1), situado na fronteira da região ótima (FIGURA 11), evidenciado por Zhao *et al.* (2012), como um possível ponto inicial para a *quadtree.* Desta forma, recomenda-se, após essa análise de viabilidade, que se compare o desempenho do método *grid-quadtree* iniciado por um ponto fixo com o por pontos aleatórios.

Por fim, apesar dos resultados do *grid-quadtree* estarem de acordo com os evidenciados na literatura, sugere-se uma análise mais aprofundada do tipo de kernel a ser aplicado na classificação das bases de dados *Liver Disorders* e *Vowel*, que apresentaram menor acurácia na previsão dos dados de teste. Nesse contexto, também se recomenda a adoção da *quadtree* em métodos que visem selecionar parâmetros para o SVC usando outras funções kernels além da gaussiana.

REFERÊNCIAS

AKAY, M. F. Support vector machines combined with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, 16, p. 3240 – 3247, 2009.

ALES, V. T. **O algoritmo *Sequential Minimal Optimisation* para resolução do problema de *Support Vector Machine***: uma técnica para reconhecimento de padrões. 150 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia) – Setores de Tecnologia e Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2008.

ALES, V. T; GEVERT, V. G; CARNIERI, C; SILVA, A. C. L. da. Análise de crédito bancário utilizando o algoritmo Sequential Minimal Optimization. IN: XLI SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL -SBPO. **Anais.** p. 2242 – 2253, 2009.

AN, H; YU, S. Well-balanced shallow water flow simulation on quadtree cut cell grids. **Advances in water resources**, 39, p. 60 – 70, 2012.

BAI, J; YANG, L; ZHANG, X. Parameter Optimization and Application of Support Vector Machine Based on Parallel Fish Swarm Algorithm. **Journal of Software**, v. 8. n.3, p. 673 – 679, 2013.

BELTRAMI, M. **Precificação de opções sobre ações por modelos de *Support Vector Regression.*** 124 f. Dissertação (Mestrado em Método Numéricos em Engenharia) – Setores de Tecnologia e Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2009.

BELTRAMI, M; SILVA, A. C. L. da. Proposta de técnica *grid-quadtree* para seleção de parâmetros do *Support Vector Classification* (SVC). In: XV SEPROSUL – SIMPOSIO DE INGENIERÍA DE LA PRODUCCIÓN SUDAMERICANO. **Anais.** 2015, Sorocaba.

BOSER, B. E; GUYON, I. M; VAPNIK, V. N. A Training Algorithm for Optimal Margin Classifiers. In: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 5., 1992, Pittsburg. **Proceedings of the 5th Annual Workshop on Computational Learning Theory**, p. 144 – 152, 1992.

CARPENTER, G.A; GROSSBERG, S; REYNOLDS, J. H. ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. **Neural Networks**, 4, p. 565 – 588, 1991.

CARVALHO, B. P. R. O estado da arte em métodos para reconhecimento de padrões: Support Vector Machine. In: CONGRESSO NACIONAL DA TECNOLOGIA DE INFORMAÇÃO E COMUNICAÇÃO. **Anais**. Belo Horizonte, 2005.

CHANG, C. C; LIN, C. J. LIBSVM: A library for Support Vector Machines. **ACM Transactions on Intelligent Systems and Technology**, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

CHAPELLE, O; VAPNIK, V; BOUSQUET, O; MUKHERJEE, S. Choosing Multiples Parameters for Support Vector Machines. **Machine Learning**, 46, p. 131- 159, 2002.

CHAPELLE, O; VAPNIK, V. Model Selection for Support Vector Machines. **Advances in Neural Information Processing Systems**, 1999.

CHAVES, A. da C. F. **Extração de regras Fuzzy para máquinas de vetores suporte (SVM) para classificação em múltiplas classes**. 225 f. Tese (Doutorado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2006. Disponível em: < http://www.maxwell.lambda.ele.puc-rio.br> Acesso em: 28/11/2008.

CHE, J. X. Support vector regression based on optimal training subset and adaptive particle swarm optimization algorithm. **Applied Soft Computing**, 13, p. 3473 – 3481, 2013.

CHERKASSKY, V; MA, Y. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. **Neural networks: the official journal of the International Neural Network**, v. 17, p. 113 – 126, 2004.

CORMEN, T. H; LEISERSON, C. E; RIVEST, R. L; STEIN, C. **Algoritmos: teoria e prática.** 3. ed. Rio de Janeiro: Elsevier, 2012.

CORTES, C; VAPNIK, V. Support – Vector Networks. **Machine Learning**, 20, 273 – 297, 1995.

CRISTIANINI, N; SHAWE-TAYLOR, J. **An Introduction to Support Vector Machines and other kernel-based learning methods**. 10th ed. United Kingdom: Cambridge University Press, 2006.

DEVOS, O; RUCKEBUSCH, C; DURAND, A; DUPONCHEL, L. HUVENNE, J. P. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. **Chemometrics and Intelligent Laboratory Systems**, 96, p. 27 – 33, 2009.

FINKEL, R. A.; BENTLEY, J. L. Quad Trees: A Data Structure for Retrieval on Composite Keys. **Acta Informatica 4,** 1 – 9, Springer- Verlag, 1974.

FRANCISQUETTI, E. P. **Estudo de quadtrees para uso de dinâmica de fluidos computacional.** 94 f. Dissertação (Mestrado em Matemática Aplicada) – Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

FRIEDMAN, J. **Another approach to polychotomous classification**. Technical report, Department of Statistics, Stanford University, 1996.

GOMES, T. A. F; PRUDÊNCIO, R. B. C; SOARES, C; ROSSI, A. L. D; CARVALHO, A. Combining meta-learning and search tecniques to select parameters for support vector machines. **Neurocomputing** 75, p. 3 – 13, 2012.

GREAVES, D. M; BORTHWICK, A. G. L Hierarchical tree-based finite element mesh generation. **International Journal for Numerical Methods is Engineering**, 45, p. 447 – 471, 1999.

GUO, Y; YANG, M; XIAO, D. The selection method for hyper-parameters of support vector classification by adaptive chaotic cultural algorithm. **International Journal of Intelligent Computing and Cybernetics**, v.3, n.3, p. 449 – 462, 2010.

GUYON, I; BOSER, B; VAPNIK, V. Automatic Capacity Tuning of Very Large VC-dimension Classifiers. **Advances in Neural Information Processing Systems 5**, p. 147 – 155, 1993.

HO, T. K; KLEINBERG, E. M. Building projectable classifiers of arbitrary complexity. **Proceedings of the 13th International Conference on Pattern Recognition**, p. 880 – 885, Vienna, Austria, August 1996.

HOSAKA, T; KOBAYASHI, T; OTSU, N. Image matting based on local color discrimination by SVM. **Pattern Recognition Letters**, 30, p. 1253- 1263, 2009.

HSU, C. W; CHANG, C. C; LIN, C. J. **A Practical Guide to Support Vector Classification***.* Technical report, Department of Computer Science, National Taiwan University, 2010. Disponível em: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Acesso em: 16/11/2015.

HSU, C. W; LIN, C. J. A Comparison of Methods for Multi-class Support Vector Machines. **IEEE Transactions on Neural Networks,** 13, p.415 – 425, 2002.

HUANG, C. L; DUN, J. F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization. **Applied Computing**, 8, p. 1381- 1391, 2008.

HUANG, C. L; WANG, C. J. A GA- based feature selection and parameters optimization for support vector machines. **Expert Systems with Applications**, 31, p.231 – 240, 2006,

JIANG, H; ZOU, L. A Hybrid PSO-SA Optimizing Approach for SVM Models in Classification. **International Journal of Biomathematics**, v. 8, n. 5, 2013.

JIANG, P; MISSOUM, S; CHEN, Z. Optimal SVM parameter selection for non-separable and unbalanced datasets. **Structural Multidisciplinary Optimization**, 50, p. 523 – 535, 2014.

JUFENG, J; LI, H; LI, P. Combined fabric defects detection approach and quadtree decomposition. **Journal of Industrial Textiles**, 41 (4), p. 331 – 344, 2011.

KAPP, M.N; SABOURIN, R; MAUPIN, P. A dynamic model selection strategy for support vector machine classifiers. **Applied Soft Computing***,* v.12, p. 2550- 2565, 2012.

KEERTHI, S. S; LIN, C. J. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. **Neural Computation**, v.15, p. 1166- 1189, 2003.

KNERR, S; PERSONNAZ, L; DREYFUS, G. Single-layer learning revisited: a stepwise procedure for building and training a neural netwoek. **Neurocomputing: Algorithms, Architectures and Applications**, Springer – Verlag, 1990.

KREBER, U. Pairwise classification and support vector machines. In SCHÖLKOPF, B; BURGES, C. J. C; SMOLA, A. J. (Ed.). **Advances in Kernel Methods – Support Vector Learning,** Cambridge: MIT Press, 1998.

LAVRENKO, V. **Introductory Applied Machine Learning: Generalization, Overfitting, Evaluation.** University of Edinburgh. School of Informatics, 2014. Disponível em: < http://homepages.inf.ed.ac.uk/vlavrenk/>. Acesso em: 29/01/2016.

LEBRUN, G; CHARRIER, C; LEZORAY, O; CARDOT, H. Tabu search model selection for SVM. **International Journal of Neural Systems**. v.18, n. 1, p. 19 – 31, 2008.

LESSMANN, S; STAHLBOCK, R; CRONE, S. F. Genetic Algorithms for Support Vector Machine Model Selection. IN: 2006 INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS. **Proceeding**, 2006.

LI, C. H; HO, H. H; LIU, Y. L; LIN, C. T; KUO, B. C; TAUR, J. S. An Automatic Method for Selecting the Parameter of the Normalized Kernel Function to Support Vector Machines. **Journal of Information Science and Engineering**, 28, p. 1 – 15, 2012.

LICHMAN, M. **UCI Machine Learning Repository.** Irvine, CA: University of California, School of Information and Computer Science, 2013. Disponível em: http://archive.ics.uci.edu/ml

LIMA, C. A de M. **Comitê de máquinas:** uma abordagem unificada empregando máquinas de vetores suporte. 342 f. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2004.

LIN, H. T; LIN, C. J. **A study on sigmoid kernels for SVM and the training of PSD kernels by SMO-type methods.**Technical report, Department of Computer Science, National Taiwan University, 2003. Disponível em: http://www.csie.ntu.edu.tw/~htlin/paper/doc/tanh.pdf. Acesso em: 31/01/2016.

LIN, K. C; HUANG, Y. H; HUNG, J. C; LIN, Y. T. Feature Selection and Parameter Optimization of Support Vector Machines Based on Modified Cat Swarm Optimization. **International Journal of Distributed Sensor Networks**, 2015.

LIN, S. W; LEE. Z. J; CHEN, S. C; TSENG, T. Y. Parameter determination of support vector machine and feature selection using simulated annealing approach. **Applied Soft Computing**, 8, p. 1505 – 1512, 2008a.

LIN, S. W; YING, K. C; CHEN, S. C; LEE, Z. J. Particle swarm optimization for parameter determination and feature selection of support vector machines. **Expert Systems with Applications**, 35, p. 1817 – 1824, 2008b.

LIU, Z; XU, H. Kernel Parameter Selection for Support Vector Machine Classification. **Journal of Algorithms & Computational Technology**, v. 8, n. 2, p.163 – 177, 2013.

MICHIE, D; SPIEGELHALTER, D. J; TAYLOR, C. C. **Machine Learning Neural and Statistical Classification**, 1994. Disponível em: http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf

MIRANDA, P. B. C; PRUDENCIO, R. B. C; CARVALHO, A. P. L. F de; SOARES, C. A hybrid meta-learning architecture for multi-objective optimization of SVM parameters. **Neurocomputing** 143, p. 27 – 43, 2014.

MOCELLIN, S; AMBROSI, A; MONTESCO, M. C; FOLETTO, M; ZAVAGNO, G; NITTI, D; LISE, M; ROSSI, C. R. Support Vector Machine Learning Model for the Prediction of Sentinel Node Status in Patients with Cutaneous Melanoma. **Annals of Surgical Oncology**. New York, v.13, n. 8, p. 1113 – 1122, 2006.

MOORE, D. The Cost of Balancing Generalized Quadtrees. **Proceedings of the 3rd Symposium on Solid Modeling and Applications**, 1995.

MUHSIN, Z. F; REHMAN, A; ALTAMEEM, A; SABA, T; UDDIN, M. Improved quadtree image segmentation approach to region information. **The Imaging Science Journal**, v. 62, n. 1, p.56 – 62, 2014.

ORTIZ-GARCIA, E. G. SALCEDO-SANZ, S; PÉREZ-BELLIDO, A. M; PORTILLA-FIGUERAS, J. A. Improving the training time of support vector regression algorithms through novel hyper-parameters search space reductions. **Neurocomputing**, 72, p. 3683 – 3691, 2009.

OSUNA, E.; FREUND, R.; GIROSI, F. Training Support Vector Machines: an Application to Face Detection. **IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. p. 130-136, 1997.

ÖZGÜNDÜZ, E.; SENTÜRK, T.; KARSLIGIL, E. Off-Line Signature Verification and Recognition by Support Vector Machine. In: 13th EUROPEAN SIGNAL PROCESSING CONFERENCE, Turkey. **Proceedings,** 2005.

PAIVA NETO, A. Quadtrees Balanceadas. Disponível em: http://w3.impa.br/~apneto/quadtrees/quadweb/balance.html. Acesso em: 15/10/2015.

PANG, H; DONG, W. D; XU, Z. H; FENG, H. J; LI, Q; CHEN, Y. T. Novel linear search for support vector machine parameter selection. **Journal of Zhejiang University – Science C (Computers & Electronics)**, v.12, p. 885- 896, 2011.

PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. In SCHÖLKOPF, B; BURGES, C. J. C; SMOLA, A. J. (Ed.). **Advances in Kernel Methods – Support Vector Learning,** Cambridge: MIT Press, 1998.

POPINET, S. Adaptive modelling of long-distance wave propagation and fine-scale flooding during the Tohoku tsunami. **Natural Hazards and Earth System Sciences**, 12, p. 1213 – 1227, 2012.

POPINET, S. Quadtree-adaptive tsunami modelling. **Ocean Dynamics**, 61, p. 1261 - 1285, 2011.

RÄTSCH, G. **IDA Benchmark Repository**, 1999. Disponível em: <http://www.raetschlab.org/Members/raetsch/benchmark>.

SAMET, H. Hierarchical spatial data structures. **Design and Implementation of Large Spatial Databases**, p. 191 – 212, 1990.

SAMET, H. Neighbor Finding in Quadtrees. **Proceedings – Institute of Electrical and Electronics Engineers Computer Society Conference on Pattern Recognition and Image Processing**, p. 68 – 74, IEEE Computer Society Press, 1981.

SAMET, H. Neighbor finding techniques for images represented by quadtrees. **Computer Graphics and Image Processing**, v.18, Issue 1, p. 37 – 57, 1982.

SAMET, H. **The Design and Analysis of Spatial Data Structures**. USA: Addison – Wesley Publishing Company, 1994.

SAMET, H. The Quadtree and Related Hierarchical Data Structure. **ACM Computing Surveys**, v.16, Issue 2, p.187 – 260, 1984.

TSAI, C. C; HOU, T. H; POPINET, S; CHAO, Y. Y. Predictions of waves generated by tropical cyclones with a quadtree adaptive model. **Coastal Engineering**, 77, p.108 -119, 2013.

VALIENTE, G. **Algorithms on Trees and Graphs.** Berlin; Heidelberg: Springer-Verlag, 2010.

VAPNIK, V. N. An Overview of Statistical Learning Theory. **IEEE Transactions on Neural Networks,** v.10, n.5, p. 988 – 999, 1999.

VAPNIK, V. N. **Statistical Learning Theory**. Wiley, NY, 1998.

VAPNIK, V. N. **The nature of statistical learning theory**. Springer – Verlag, New York, 1995.

VAPNIK, V; CHERVONENKIS, A. A note on one class of perceptrons. **Automation and Remote Control**, 25, 1964.

VAPNIK, V; LERNER, A. Pattern recognition using generalized portrait method. **Automation and Remote Control,** 24, p. 774 – 780, 1963.

VAREWYCK, M; MARTENS J. P. A Practical Approach to Model Selection for Support Vector Machines With a Gaussian Kernel. **IEEE Transactions on Systems Man. and Cybernectis – Part B: Cybernectis**, v. 41, n. 2, 2011.

WANG, X; HUANG, F; CHENG, Y. Super-parameter selection for Gaussian-Kernel SVM based on outlier-resisting. **Measurement***,* 58, p.147 – 153, 2014.

WANG, X; YANG, C; QIN, B; GUI, W. Parameter selection of support vector regression based on hybrid optimization algorithm and its application. **Journal of Control Theory and Applications**, 4, p. 371 -376, 2005.

YANG, D; REINDL, T. Solar irradiance monitoring network design using the variance quadtree algorithm. **Renewables: Wind, Water, and Solar**, 2:1, 2015.

YAO, Y; MARCIALIS, G. L; PONTIL, M; FRASCONI, P; ROLI, F. A New Machine Learning Approach to Fingerprint Classification. **Advances in Artificial Intelligence**, v.2175. p. 57-63, 2001.

YUAN, F. C. Parameters Optimization Using Genetic Algorithms in Support Vector Regression for Sales Volume Forecasting. **Applied Mathematics**, 3, p. 1480 – 1486, 2012.

YUEN, C. H; LUI, O. Y; WONG, K. W. Hybrid fractal image coding with quadtree-based progressive structure. **Journal of Visual Communication & Image Representation**, 24, p. 1328 – 1341, 2013.

ZHANG, X; CHEN, X; HE, Z. An ACO-based algorithm for parameter optimization of support vector machines. **Expert Systems with Applications**, 37, p. 6618 – 6628, 2010.

ZHANG, X; CHEN, W; WANG, B; CHEN, X. Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization. **Neurocomputing**, 167, p.260 – 279, 2015.

ZHAO, M; FU, C; JI, L; TANG, K; ZHOU, M. Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. **Expert Systems with Applications**, 38, p. 5197 – 5204, 2011.

ZHAO, M; REN, J. JI, L; FU, C; LI, J; ZHOU, M. Parameter selection of support vector machines and genetic algorithm based on change area search. **Neural Computing & Applications**, 21: 1 – 8, 2012.

APÊNDICE

No apêndice deste trabalho apresentam-se os resultados referentes às bases de dados *Ionosphere, Breast Cancer, Diabetes, Circle and Square, Iris, Svmguide* 2, *Vehicle* e *Segment*, estudadas no grupo 1 (QUADRO 20). Dentre esses estão: os parâmetros encontrados pela busca por *grid* (BG) e pelo *grid-quadtree* (GQ), o número de operações realizadas por ambos os métodos e a quantidade de acertos (QA) e a acurácia atingida pelo SVC-BG e SVC-GQ, para cada um dos conjuntos avaliados (A1 a A30). Tais resultados encontram-se nas TABELAS 25 a 48 e nos GRÁFICOS 2 a 9, cujas interpretações são realizadas conforme explicado na seção 5.1.

Ao final deste apêndice constam também os QUADROS 28 a 30, que relacionam respectivamente as referências bibliográficas desta tese com as bases de dados abordadas no estudo estatístico da seção 4.6.4 e nos grupos 1 e 2 de dados.

TABELA 25 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *IONOSPHERE*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG/GQ	22,6274	0,1250	94,6619	71	25,2669	2
A2	BG	8,0000	0,0625	95,7295	62	22,0641	16
	GQ	16,0000	0,0526	95,7295	55	19,5730	12
A3	BG	16,0000	0,0884	95,7295	61	21,7082	8
	GQ	24,6754	0,0537	95,7295	57	20,2847	9
A4	BG/GQ	2,0000	0,1250	95,7295	92	32,7402	22
A5	BG	1,4142	0,3536	95,0178	129	45,9075	12
	GQ	2,0000	0,1768	93,9502	112	39,8577	17
A6	BG	5,6569	0,3536	95,7295	127	45,1957	3
	GQ	5,1874	0,3536	95,7295	127	45,1957	4
A7	BG	4,0000	0,1768	94,6619	108	38,4342	10
	GQ	1,9152	0,1393	94,6619	98	34,8755	22
A8	BG	8,0000	0,2500	97,1530	115	40,9253	3
	GQ	6,8745	0,2668	96,7972	116	41,2811	3
A9	BG/GQ	16,0000	0,0884	96,0854	53	18,8612	5
A10	BG	5,6569	0,3536	94,3060	130	46,2633	4
	GQ	4,1771	0,3692	94,3060	131	46,6192	5
A11	BG	16,0000	0,0442	96,4413	56	19,9288	11
	GQ	20,3048	0,0423	96,7972	56	19,9288	10
A12	BG	5,6569	0,1768	95,0178	103	36,6548	8
	GQ	7,1788	0,1393	95,0178	79	28,1139	8

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A13	BG/GQ	22,6274	0,0625	95,3737	53	18,8612	7
A14	BG	5,6569	0,2500	95,7295	119	42,3488	7
	GQ	8,0000	0,2668	95,7295	118	41,9929	5
A15	BG/GQ	64,0000	0,0313	94,6619	47	16,7260	6
A16	BG	8,0000	0,0442	96,0854	57	20,2847	18
	GQ	9,9349	0,0442	96,0854	56	19,9288	15
A17	BG	16,0000	0,1250	95,7295	64	22,7758	5
	GQ	26,3321	0,0884	94,6619	55	19,5730	5
A18	BG/GQ	11,3137	0,0442	95,7295	59	20,9964	17
A19	BG	2,8284	0,0625	94,6619	73	25,9786	34
	GQ	4,0000	0,0492	94,6619	71	25,2669	32
A20	BG	8,0000	0,0442	95,0178	60	21,3523	21
	GQ	69,7925	0,0326	94,3060	47	16,7260	6
A21	BG/GQ	16,0000	0,3536	95,0178	130	46,2633	1
A22	BG/GQ	4,0000	0,0884	95,7295	67	23,8434	18
A23	BG	16,0000	0,1250	95,0178	71	25,2669	6
	GQ	18,2206	0,1098	95,0178	69	24,5552	6
A24	BG	11,3137	0,0625	96,0854	60	21,3523	10
	GQ	41,4989	0,0442	95,7295	52	18,5053	6
A25	BG	45,2548	0,3536	96,7972	123	43,7722	0
	GQ	197,4030	0,3536	96,7972	123	43,7722	0
A26	BG	11,3137	0,0313	96,4413	59	20,9964	19
	GQ	30,6433	0,0442	96,0854	50	17,7936	7
A27	BG/GQ	22,6274	0,3536	96,7972	121	43,0605	0
A28	BG	11,3137	0,3536	96,0854	126	44,8399	2
	GQ	8,0000	0,1768	95,7295	103	36,6548	6
A29	BG	1,0000	0,1768	95,7295	115	40,9253	36
	GQ	3,7483	0,0727	95,0178	69	24,5552	24
A30	BG	22,6274	0,1768	95,0178	99	35,2313	1
	GQ	23,6292	0,2013	95,0178	107	38,0783	1

FONTE: A autora (2016).

TABELA 26 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *IONOSPHERE*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	130	100	11,9376	-	88,0624
A2	1089	164	100	15,0597	-	84,9403
A3	1089	226	100	20,7530	-	79,2470
A4	1089	132	100	12,1212	-	87,8788
A5	1089	136	100	12,4885	-	87,5115
A6	1089	164	100	15,0597	-	84,9403
A7	1089	155	100	14,2332	-	85,7668
A8	1089	135	100	12,3967	-	87,6033
A9	1089	133	100	12,2130	-	87,7870
A10	1089	218	100	20,0184	-	79,9816
A11	1089	203	100	18,6410	-	81,3591
A12	1089	136	100	12,4885	-	87,5115
A13	1089	132	100	12,1212	-	87,8788
A14	1089	158	100	14,5087	-	85,4913
A15	1089	129	100	11,8457	-	88,1543
A16	1089	134	100	12,3049	-	87,6951
A17	1089	134	100	12,3049	-	87,6951
A18	1089	132	100	12,1212	-	87,8788
A19	1089	232	100	21,3040	-	78,6961
A20	1089	140	100	12,8558	-	87,1442
A21	1089	133	100	12,2130	-	87,7870
A22	1089	221	100	20,2939	-	79,7062
A23	1089	146	100	13,4068	-	86,5932
A24	1089	142	100	13,0395	-	86,9605
A25	1089	229	100	21,0285	-	78,9715
A26	1089	159	100	14,6006	-	85,3995
A27	1089	211	100	19,3756	-	80,6244
A28	1089	135	100	12,3967	-	87,6033
A29	1089	133	100	12,2130	-	87,7870
A30	1089	228	100	20,9366	-	79,0634

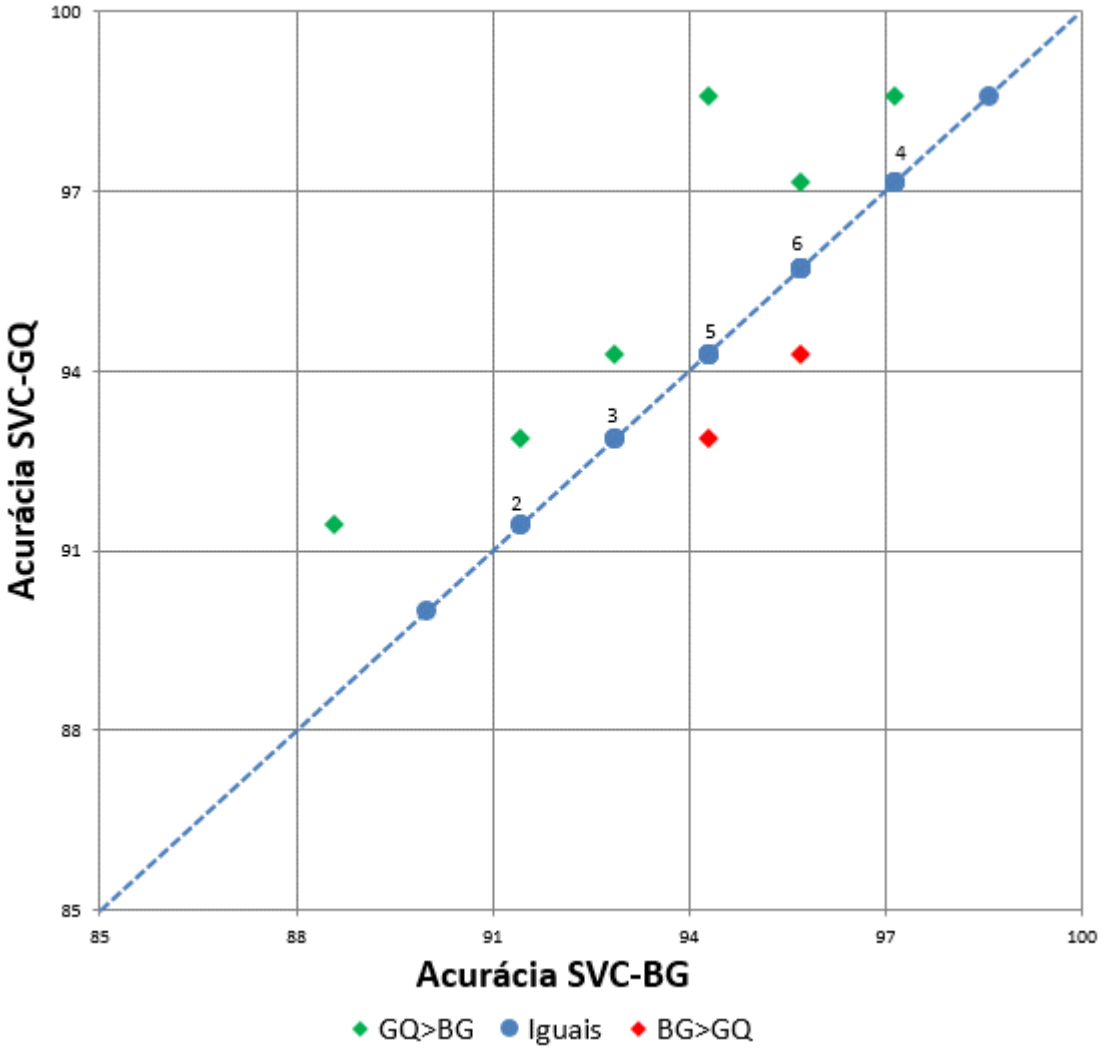
FONTE: A autora (2016)

TABELA 27 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *IONOSPHERE*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	66/70	94,2857	66/70	94,2857
A2	67/70	95,7143	66/70	94,2857
A3	67/70	95,7143	67/70	95,7143
A4	65/70	92,8571	65/70	92,8571
A5	68/70	97,1429	69/70	98,5714
A6	69/70	98,5714	69/70	98,5714
A7	67/70	95,7143	67/70	95,7143
A8	66/70	94,2857	66/70	94,2857
A9	64/70	91,4286	64/70	91,4286
A10	68/70	97,1429	68/70	97,1429
A11	65/70	92,8571	65/70	92,8571
A12	68/70	97,1429	68/70	97,1429
A13	66/70	94,2857	66/70	94,2857
A14	64/70	91,4286	64/70	91,4286
A15	66/70	94,2857	66/70	94,2857
A16	64/70	91,4286	65/70	92,8571
A17	67/70	95,7143	67/70	95,7143
A18	67/70	95,7143	67/70	95,7143
A19	67/70	95,7143	68/70	97,1429
A20	65/70	92,8571	66/70	94,2857
A21	68/70	97,1429	68/70	97,1429
A22	65/70	92,8571	65/70	92,8571
A23	68/70	97,1429	68/70	97,1429
A24	67/70	95,7143	67/70	95,7143
A25	67/70	95,7143	67/70	95,7143
A26	62/70	88,5714	64/70	91,4286
A27	63/70	90,0000	63/70	90,0000
A28	66/70	94,2857	65/70	92,8571
A29	66/70	94,2857	69/70	98,5714
A30	66/70	94,2857	66/70	94,2857

FONTE: A autora (2016).

GRÁFICO 2 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *IONOSPHERE*



FONTE: A autora (2016).

TABELA 28 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *BREAST CANCER*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG	0,7071	0,1250	97,0696	54	9,8901	39
	GQ	8,7241	0,1098	96,8864	43	7,8755	20
A2	BG	90,5097	0,0221	97,8022	33	6,0440	17
	GQ	76,1093	0,0241	97,8022	33	6,0440	17
A3	BG	1,4142	0,0884	97,8022	44	8,0586	33
	GQ	2,0000	0,0579	97,8022	42	7,6923	34
A4	BG	2,0000	0,3536	97,2527	76	13,9194	19
	GQ	6,7272	0,7071	97,2527	98	17,9487	3
A5	BG	0,5000	0,0884	97,2527	58	10,6227	49
	GQ	0,5000	0,0811	97,2527	60	10,9890	50
A6	BG	5,6569	0,0055	96,8864	61	11,1722	56
	GQ	22,6274	0,0442	96,7033	44	8,0586	25
A7	BG	8,0000	0,0313	97,4359	39	7,1429	31
	GQ	11,3137	0,0280	97,4359	39	7,1429	29
A8	BG	64,0000	0,0078	97,4359	37	6,7766	27
	GQ	64,0000	0,0075	97,4359	37	6,7766	27
A9	BG	256,0000	0,0055	96,8864	38	6,9597	27
	GQ	256,0000	0,0110	96,8864	39	7,1429	24
A10	BG	90,5097	0,0039	97,0696	40	7,3260	33
	GQ	1,4142	0,1768	96,8864	57	10,4396	32
A11	BG	5,6569	0,0625	97,2527	41	7,5092	26
	GQ	5,6569	0,0653	97,2527	41	7,5092	26
A12	BG	32,0000	0,0055	97,2527	42	7,6923	35
	GQ	36,4412	0,0066	97,2527	40	7,3260	34
A13	BG	2,0000	0,0442	97,2527	47	8,6081	39
	GQ	7,6608	0,0313	97,0696	40	7,3260	29
A14	BG	16,0000	0,0442	97,6190	35	6,4103	20
	GQ	16,7084	0,0442	97,6190	35	6,4103	19
A15	BG	16,0000	0,0156	96,8864	42	7,6923	32
	GQ	26,3321	0,0110	96,5201	41	7,5092	32
A16	BG	64,0000	0,0221	97,8022	35	6,4103	18
	GQ	90,5097	0,0186	97,8022	35	6,4103	18
A17	BG	0,5000	0,0313	96,8864	71	13,0037	66
	GQ	22,6274	0,0039	96,5201	44	8,0586	38
A18	BG	22,6274	0,0313	97,8022	32	5,8608	20
	GQ	19,0273	0,0333	97,8022	32	5,8608	21
A19	BG	5,6569	0,0625	97,0696	46	8,4249	32
	GQ	69,7925	0,0178	96,8864	42	7,6923	27

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A20	BG	1,4142	0,5000	97,2527	93	17,0330	18
	GQ	1,9571	0,1007	97,0696	45	8,2418	30
A21	BG/GQ	2,8284	0,2500	97,2527	56	10,2564	18
A22	BG/GQ	0,3536	0,0884	97,0696	65	11,9048	55
A23	BG	45,2548	0,0110	97,8022	33	6,0440	22
	GQ	15,3217	0,0423	97,8022	35	6,4103	20
A24	BG	1,0000	0,0884	97,0696	51	9,3407	41
	GQ	22,6274	0,0039	96,8864	46	8,4249	40
A25	BG/GQ	22,6274	0,0110	97,2527	40	7,3260	31
A26	BG	2,0000	0,0625	97,4359	42	7,6923	33
	GQ	76,1093	0,0039	97,2527	35	6,4103	28
A27	BG	1,0000	0,3536	97,8022	67	12,2711	21
	GQ	96,5865	0,0078	97,4359	27	4,9451	18
A28	BG	2,0000	0,1250	97,0696	50	9,1575	29
	GQ	3,7483	0,0884	97,0696	45	8,2418	28
A29	BG	0,3536	0,0625	97,0696	67	12,2711	59
	GQ	128,0000	0,0039	96,8864	35	6,4103	25
A30	BG	0,7071	0,0884	97,4359	51	9,3407	41
	GQ	0,8140	0,0811	97,4359	50	9,1575	40

FONTE: A autora (2016).

TABELA 29 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *BREAST CANCER*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	156	100	14,3251	-	85,6749
A2	1089	163	100	14,9679	-	85,0321
A3	1089	240	100	22,0386	-	77,9614
A4	1089	202	100	18,5491	-	81,4509
A5	1089	130	100	11,9376	-	88,0624
A6	1089	132	100	12,1212	-	87,8788
A7	1089	183	100	16,8044	-	83,1956
A8	1089	160	100	14,6924	-	85,3076
A9	1089	228	100	20,9366	-	79,0634
A10	1089	135	100	12,3967	-	87,6033
A11	1089	166	100	15,2433	-	84,7567
A12	1089	162	100	14,8760	-	85,1240
A13	1089	159	100	14,6006	-	85,3995
A14	1089	139	100	12,7640	-	87,2360
A15	1089	129	100	11,8457	-	88,1543
A16	1089	195	100	17,9063	-	82,0937
A17	1089	118	100	10,8356	-	89,1644
A18	1089	227	100	20,8448	-	79,1552
A19	1089	169	100	15,5188	-	84,4812
A20	1089	131	100	12,0294	-	87,9706
A21	1089	159	100	14,6006	-	85,3995
A22	1089	135	100	12,3967	-	87,6033
A23	1089	165	100	15,1515	-	84,8485
A24	1089	98	100	8,9991	-	91,0009
A25	1089	131	100	12,0294	-	87,9706
A26	1089	177	100	16,2534	-	83,7466
A27	1089	133	100	12,2130	-	87,7870
A28	1089	575	100	52,8007	-	47,1993
A29	1089	200	100	18,3655	-	81,6345
A30	1089	201	100	18,4573	-	81,5427

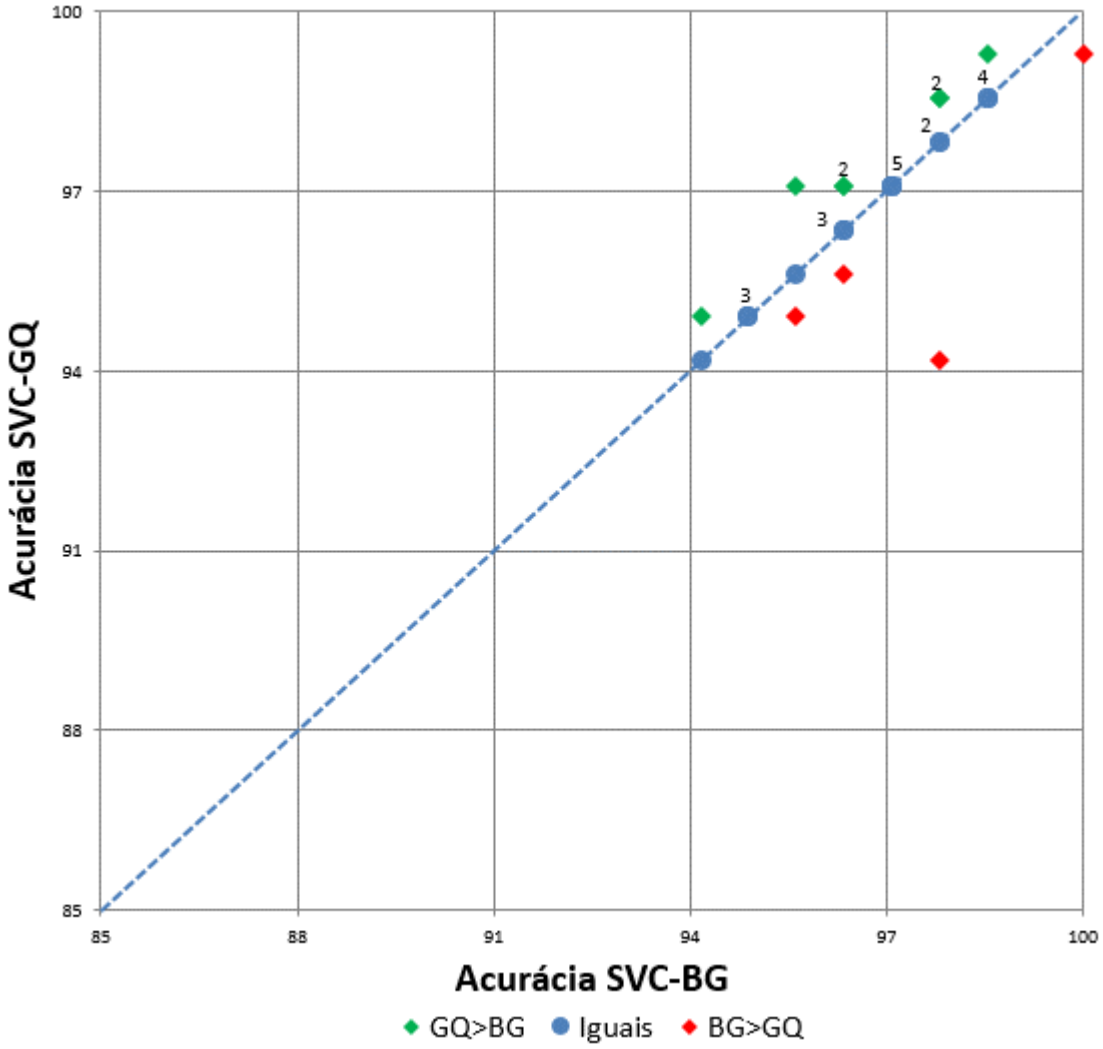
FONTE: A autora (2016).

TABELA 30 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *BREAST CANCER*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	133/137	97,0803	133/137	97,0803
A2	129/137	94,1606	129/137	94,1606
A3	132/137	96,3504	131/137	95,6204
A4	134/137	97,8102	129/137	94,1606
A5	132/137	96,3504	132/137	96,3504
A6	135/137	98,5401	136/137	99,2701
A7	133/137	97,0803	133/137	97,0803
A8	133/137	97,0803	133/137	97,0803
A9	134/137	97,8102	135/137	98,5401
A10	135/137	98,5401	135/137	98,5401
A11	133/137	97,0803	133/137	97,0803
A12	134/137	97,8102	134/137	97,8102
A13	134/137	97,8102	134/137	97,8102
A14	130/137	94,8905	130/137	94,8905
A15	135/137	98,5401	135/137	98,5401
A16	130/137	94,8905	130/137	94,8905
A17	132/137	96,3504	133/137	97,0803
A18	130/137	94,8905	130/137	94,8905
A19	137/137	100,0000	136/137	99,2701
A20	131/137	95,6204	133/137	97,0803
A21	131/137	95,6204	131/137	95,6204
A22	133/137	97,0803	133/137	97,0803
A23	131/137	95,6204	130/137	94,8905
A24	134/137	97,8102	135/137	98,5401
A25	135/137	98,5401	135/137	98,5401
A26	132/137	96,3504	132/137	96,3504
A27	129/137	94,1606	130/137	94,8905
A28	135/137	98,5401	135/137	98,5401
A29	132/137	96,3504	132/137	96,3504
A30	132/137	96,3504	133/137	97,0803

FONTE: A autora (2016).

GRÁFICO 3 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *BREAST CANCER*



FONTE: A autora (2016).

TABELA 31 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *DIABETES*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG	2,0000	0,1768	78,6645	280	45,6026	261
	GQ	189,0338	0,0274	78,5016	260	42,3453	235
A2	BG	0,7071	0,0884	76,8730	319	51,9544	312
	GQ	256,0000	0,0085	76,0586	263	42,8339	250
A3	BG	256,0000	0,0110	75,8958	268	43,6482	253
	GQ	117,3765	0,0156	75,8958	268	43,6482	254
A4	BG	2,8284	0,3536	78,8274	257	41,8567	227
	GQ	1,3252	0,2500	78,8274	271	44,1368	254
A5	BG	8,0000	0,0313	76,3844	289	47,0684	279
	GQ	6,7272	0,0442	76,2215	286	46,5798	275
A6	BG	32,0000	0,0078	78,3388	276	44,9511	268
	GQ	181,0193	0,0039	78,0130	262	42,6710	253
A7	BG/GQ	16,0000	0,0313	76,8730	270	43,9739	258
A8	BG	32,0000	0,0313	77,6873	262	42,6710	249
	GQ	64,0000	0,0274	77,8502	259	42,1824	243
A9	BG	1,4142	0,0625	78,0130	297	48,3713	287
	GQ	32,0000	0,0110	77,5244	263	42,8339	253
A10	BG	181,0193	0,0156	78,1759	252	41,0423	237
	GQ	82,9977	0,0221	78,1759	254	41,3681	239
A11	BG	11,3137	0,0313	77,3616	279	45,4397	269
	GQ	128,0000	0,0078	77,1987	270	43,9739	259
A12	BG	8,0000	0,0884	78,0130	259	42,1824	242
	GQ	45,2548	0,0313	77,8502	256	41,6938	240
A13	BG	16,0000	0,0078	77,1987	293	47,7199	285
	GQ	47,2584	0,0313	77,0358	258	42,0195	243
A14	BG	2,8284	0,0442	77,8502	293	47,7199	285
	GQ	19,0273	0,0760	78,0130	256	41,6938	237
A15	BG	45,2548	0,0313	77,5244	266	43,3225	253
	GQ	1,1764	0,2786	77,1987	286	46,5798	269
A16	BG	16,0000	0,0625	80,6189	245	39,9023	228
	GQ	26,3321	0,0442	80,9446	244	39,7394	227
A17	BG	5,6569	0,0156	78,3388	301	49,0228	293
	GQ	11,3137	0,0313	77,8502	273	44,4625	261
A18	BG	5,6569	0,0884	77,0358	267	43,4853	252
	GQ	10,4877	0,0579	77,0358	264	42,9967	250
A19	BG	0,7071	0,2500	78,6645	293	47,7199	279
	GQ	0,7384	0,2726	78,8274	291	47,3941	275
A20	BG	90,5097	0,0156	78,1759	259	42,1824	246
	GQ	45,2548	0,0442	78,0130	257	41,8567	237

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A21	BG	0,5000	0,1768	77,0358	312	50,8143	302
	GQ	26,3321	0,0333	76,5472	267	43,4853	252
A22	BG	2,0000	0,0442	77,5244	306	49,8371	299
	GQ	3,6680	0,0241	77,5244	305	49,6743	298
A23	BG	90,5097	0,0442	78,9902	243	39,5765	221
	GQ	107,6347	0,0625	78,6645	239	38,9251	212
A24	BG	0,7071	2,8284	78,5016	330	53,7459	227
	GQ	18,2206	0,0712	78,3388	255	41,5309	236
A25	BG	11,3137	0,0156	78,6645	277	45,1140	268
	GQ	1,0000	0,2013	78,8274	280	45,6026	261
A26	BG	90,5097	0,0221	77,5244	264	42,9967	248
	GQ	45,2548	0,0442	77,0358	262	42,6710	242
A27	BG	16,0000	0,7071	77,3616	252	41,0423	145
	GQ	11,3137	0,1098	77,3616	255	41,5309	234
A28	BG	256,0000	0,0156	78,0130	256	41,6938	238
	GQ	145,7649	0,0221	78,0130	257	41,8567	237
A29	BG	1,4142	0,3536	77,3616	275	44,7883	250
	GQ	2,0000	0,2500	77,1987	273	44,4625	250
A30	BG	4,0000	0,0625	78,8274	269	43,8111	258
	GQ	12,3377	0,0442	78,6645	257	41,8567	245

FONTE: A autora (2016).

TABELA 32 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *DIABETES*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	148	100	13,5905	-	86,4096
A2	1089	185	100	16,9881	-	83,0119
A3	1089	198	100	18,1818	-	81,8182
A4	1089	180	100	16,5289	-	83,4711
A5	1089	129	100	11,8457	-	88,1543
A6	1089	267	100	24,5179	-	75,4821
A7	1089	132	100	12,1212	-	87,8788
A8	1089	174	100	15,9780	-	84,0220
A9	1089	129	100	11,8457	-	88,1543
A10	1089	145	100	13,3150	-	86,6850
A11	1089	234	100	21,4876	-	78,5124
A12	1089	134	100	12,3049	-	87,6951
A13	1089	210	100	19,2838	-	80,7163
A14	1089	267	100	24,5179	-	75,4821
A15	1089	133	100	12,2130	-	87,7870
A16	1089	167	100	15,3352	-	84,6648
A17	1089	133	100	12,2130	-	87,7870
A18	1089	482	100	44,2608	-	55,7392
A19	1089	133	100	12,2130	-	87,7870
A20	1089	131	100	12,0294	-	87,9706
A21	1089	174	100	15,9780	-	84,0220
A22	1089	134	100	12,3049	-	87,6951
A23	1089	151	100	13,8659	-	86,1341
A24	1089	180	100	16,5289	-	83,4711
A25	1089	128	100	11,7539	-	88,2461
A26	1089	151	100	13,8659	-	86,1341
A27	1089	243	100	22,3141	-	77,6860
A28	1089	136	100	12,4885	-	87,5115
A29	1089	139	100	12,7640	-	87,2360
A30	1089	132	100	12,1212	-	87,8788

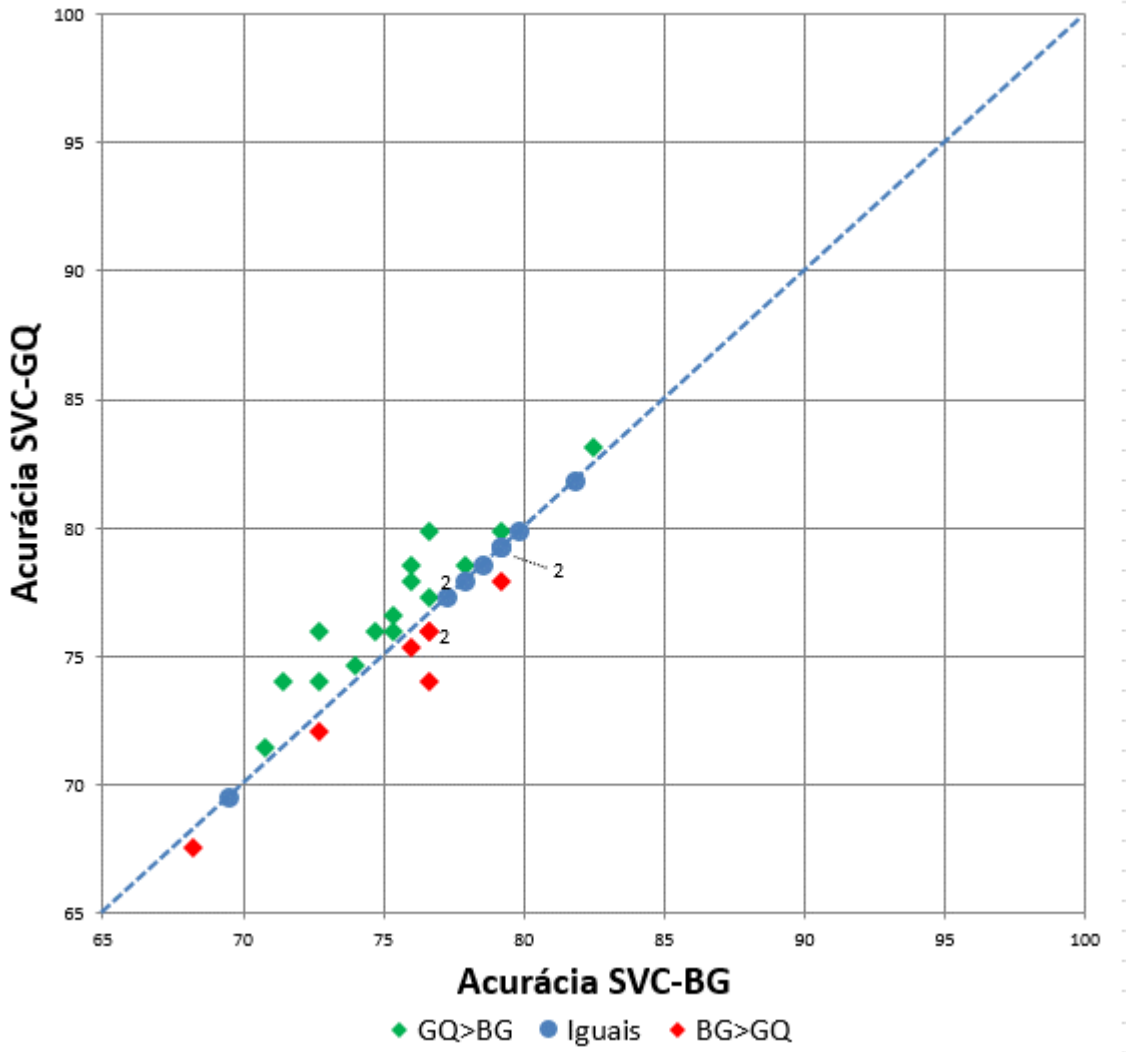
FONTE: A autora (2016).

TABELA 33 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *DIABETES*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	118/154	76,6234	114/154	74,0260
A2	118/154	76,6234	123/154	79,8701
A3	126/154	81,8182	126/154	81,8182
A4	112/154	72,7273	114/154	74,0260
A5	127/154	82,4675	128/154	83,1169
A6	116/154	75,3247	118/154	76,6234
A7	123/154	79,8701	123/154	79,8701
A8	120/154	77,9221	121/154	78,5714
A9	114/154	74,0260	115/154	74,6753
A10	116/154	75,3247	117/154	75,9740
A11	120/154	77,9221	120/154	77,9221
A12	119/154	77,2727	119/154	77,2727
A13	117/154	75,9740	121/154	78,5714
A14	118/154	76,6234	117/154	75,9740
A15	122/154	79,2208	122/154	79,2208
A16	105/154	68,1818	104/154	67,5325
A17	118/154	76,6234	119/154	77,2727
A18	122/154	79,2208	122/154	79,2208
A19	118/154	76,6234	117/154	75,9740
A20	115/154	74,6753	117/154	75,9740
A21	121/154	78,5714	121/154	78,5714
A22	122/154	79,2208	120/154	77,9221
A23	107/154	69,4805	107/154	69,4805
A24	110/154	71,4286	114/154	74,0260
A25	109/154	70,7792	110/154	71,4286
A26	117/154	75,9740	120/154	77,9221
A27	112/154	72,7273	117/154	75,9740
A28	117/154	75,9740	116/154	75,3247
A29	122/154	79,2208	123/154	79,8701
A30	112/154	72,7273	111/154	72,0779

FONTE: A autora (2016).

GRÁFICO 4 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *DIABETES*



FONTE: A autora (2016)

TABELA 34 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *CIRCLE AND SQUARE*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A1	BG	128,0000	11,3137	99,3750	41	5,1250	1
	GQ	117,3765	11,3137	99,3750	41	5,1250	1
A2	BG	256,0000	2,8284	99,5000	25	3,1250	7
	GQ	256,0000	2,2776	99,5000	25	3,1250	9
A3	BG	181,0193	2,8284	99,3750	29	3,6250	10
	GQ	173,3447	2,8284	99,3750	29	3,6250	11
A4	BG	256,0000	0,1768	99,5000	51	6,3750	44
	GQ	107,6347	0,7071	99,2500	36	4,5000	27
A5	BG	45,2548	16,0000	99,5000	60	7,5000	6
	GQ	197,4030	16,0000	99,3750	54	6,7500	1
A6	BG	32,0000	2,8284	99,6250	51	6,3750	33
	GQ	41,4989	1,6818	99,6250	48	6,0000	35
A7	BG	45,2548	0,5000	99,3750	60	7,5000	51
	GQ	239,8935	0,5000	99,2500	35	4,3750	26
A8	BG	256,0000	2,8284	99,7500	26	3,2500	6
	GQ	234,7530	4,0000	99,7500	27	3,3750	4
A9	BG/GQ	256,0000	2,0000	99,7500	23	2,8750	6
A10	BG	45,2548	1,4142	98,8750	44	5,5000	32
	GQ	256,0000	2,2776	98,7500	28	3,5000	11
A11	BG	256,0000	1,4142	99,2500	27	3,3750	11
	GQ	64,0000	2,8284	99,2500	39	4,8750	22
A12	BG	64,0000	0,1768	99,5000	83	10,3750	76
	GQ	50,4314	1,0905	99,5000	44	5,5000	33
A13	BG	256,0000	0,5000	99,6250	32	4,0000	23
	GQ	133,6670	1,3543	99,6250	32	4,0000	18
A14	BG	45,2548	0,2500	99,3750	79	9,8750	72
	GQ	66,1136	1,7754	99,2500	40	5,0000	25
A15	BG	256,0000	2,8284	99,7500	25	3,1250	5
	GQ	125,2572	4,1771	99,8750	30	3,7500	8
A16	BG	181,0193	0,1768	99,7500	59	7,3750	51
	GQ	66,1136	0,3105	99,6250	62	7,7500	54
A17	BG	256,0000	5,6569	99,0000	30	3,7500	4
	GQ	36,4412	14,0500	99,0000	58	7,2500	11
A18	BG	256,0000	0,0884	99,5000	77	9,6250	69
	GQ	173,3447	0,1098	99,5000	76	9,5000	69
A19	BG	90,5097	1,4142	99,5000	37	4,6250	23
	GQ	76,1093	1,7186	99,5000	38	4,7500	23
A20	BG	90,5097	1,4142	99,1250	39	4,8750	26
	GQ	66,8335	1,3690	99,1250	41	5,1250	29

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)	VS*Bound*
A21	BG	181,0193	0,0884	99,2500	91	11,3750	84
	GQ	256,0000	2,2776	99,3750	27	3,3750	9
A22	BG	181,0193	1,0000	99,5000	29	3,6250	17
	GQ	256,0000	0,1552	99,3750	54	6,7500	48
A23	BG	256,0000	1,0000	99,3750	28	3,5000	17
	GQ	256,0000	0,9371	99,3750	29	3,6250	17
A24	BG	128,0000	5,6569	99,5000	32	4,0000	8
	GQ	165,9955	5,1874	99,3750	31	3,8750	6
A25	BG	181,0193	1,0000	99,5000	29	3,6250	17
	GQ	173,3447	0,9170	99,3750	30	3,7500	19
A26	BG	64,0000	1,0000	99,5000	41	5,1250	30
	GQ	76,1093	1,4142	99,3750	38	4,7500	24
A27	BG	181,0193	2,0000	99,5000	30	3,7500	13
	GQ	256,0000	3,0844	99,5000	28	3,5000	7
A28	BG	128,0000	1,0000	99,3750	34	4,2500	22
	GQ	184,9831	0,5453	99,5000	34	4,2500	24
A29	BG	128,0000	1,4142	99,6250	31	3,8750	19
	GQ	109,9916	1,3543	99,6250	34	4,2500	21
A30	BG	128,0000	0,2500	99,5000	56	7,0000	49
	GQ	96,5865	0,7071	99,2500	40	5,0000	30

FONTE: A autora (2016).

TABELA 35 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *CIRCLE AND SQUARE*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	226	100	20,7530	-	79,2470
A2	1089	257	100	23,5996	-	76,4004
A3	1089	232	100	21,3040	-	78,6961
A4	1089	187	100	17,1717	-	82,8283
A5	1089	207	100	19,0083	-	80,9917
A6	1089	155	100	14,2332	-	85,7668
A7	1089	182	100	16,7126	-	83,2874
A8	1089	167	100	15,3352	-	84,6648
A9	1089	168	100	15,4270	-	84,5730
A10	1089	188	100	17,2635	-	82,7365
A11	1089	133	100	12,2130	-	87,7870
A12	1089	152	100	13,9578	-	86,0422
A13	1089	149	100	13,6823	-	86,3177
A14	1089	251	100	23,0487	-	76,9513
A15	1089	227	100	20,8448	-	79,1552
A16	1089	228	100	20,9366	-	79,0634
A17	1089	160	100	14,6924	-	85,3076
A18	1089	240	100	22,0386	-	77,9614
A19	1089	234	100	21,4876	-	78,5124
A20	1089	367	100	33,7006	-	66,2994
A21	1089	236	100	21,6713	-	78,3287
A22	1089	296	100	27,1809	-	72,8191
A23	1089	263	100	24,1506	-	75,8494
A24	1089	337	100	30,9458	-	69,0542
A25	1089	239	100	21,9467	-	78,0533
A26	1089	139	100	12,7640	-	87,2360
A27	1089	249	100	22,8650	-	77,1350
A28	1089	231	100	21,2121	-	78,7879
A29	1089	214	100	19,6511	-	80,3489
A30	1089	212	100	19,4674	-	80,5326

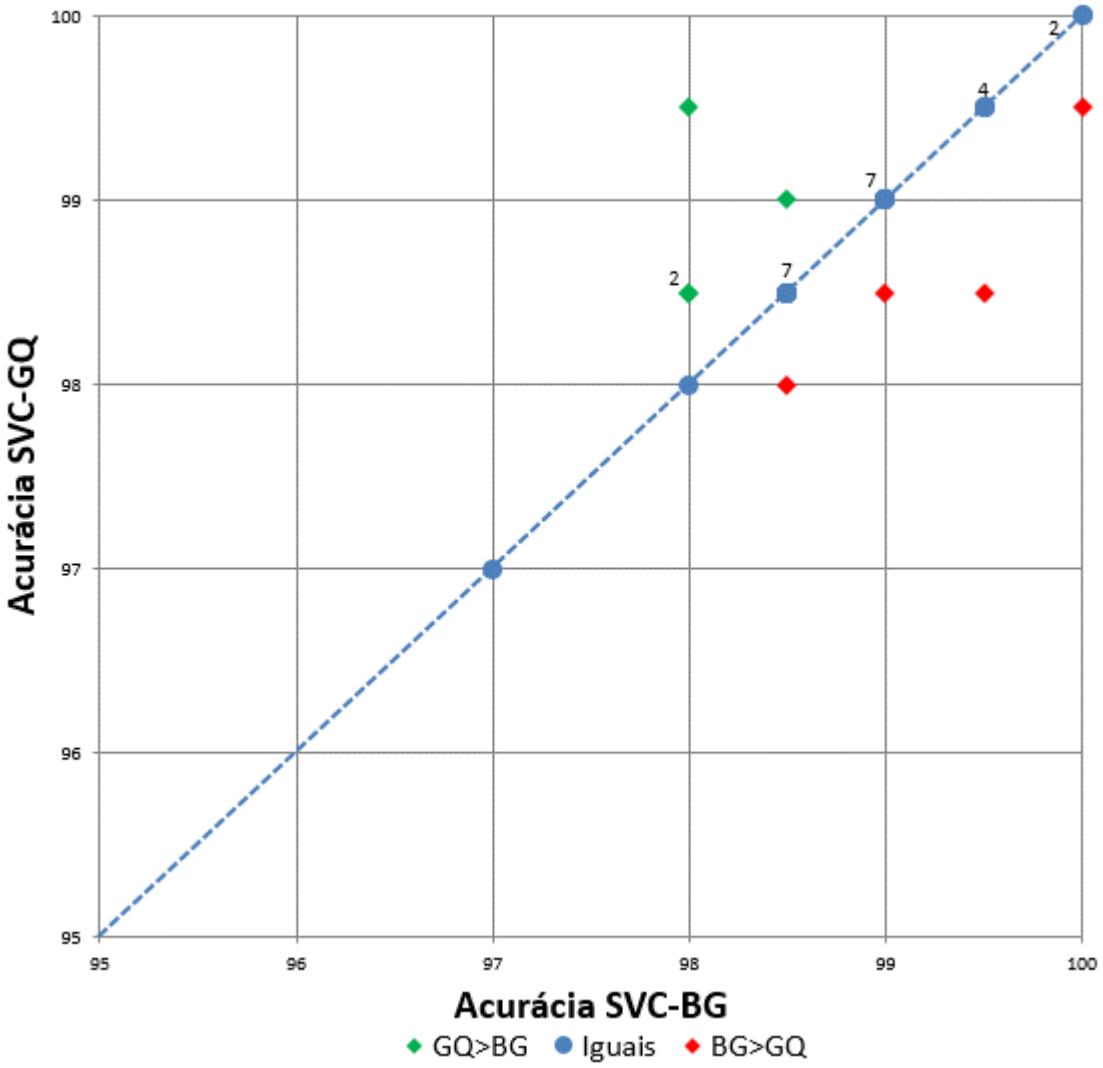
FONTE: A autora (2016).

TABELA 36 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *CIRCLE AND SQUARE*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	197/200	98,5000	197/200	98,5000
A2	196/200	98,0000	197/200	98,5000
A3	198/200	99,0000	198/200	99,0000
A4	198/200	99,0000	197/200	98,5000
A5	197/200	98,5000	197/200	98,5000
A6	200/200	100,0000	200/200	100,0000
A7	199/200	99,5000	199/200	99,5000
A8	197/200	98,5000	197/200	98,5000
A9	194/200	97,0000	194/200	97,0000
A10	200/200	100,0000	199/200	99,5000
A11	196/200	98,0000	199/200	99,5000
A12	198/200	99,0000	198/200	99,0000
A13	197/200	98,5000	198/200	99,0000
A14	199/200	99,5000	199/200	99,5000
A15	196/200	98,0000	196/200	98,0000
A16	197/200	98,5000	197/200	98,5000
A17	199/200	99,5000	199/200	99,5000
A18	198/200	99,0000	198/200	99,0000
A19	198/200	99,0000	198/200	99,0000
A20	199/200	99,5000	199/200	99,5000
A21	199/200	99,5000	197/200	98,5000
A22	196/200	98,0000	197/200	98,5000
A23	197/200	98,5000	197/200	98,5000
A24	200/200	100,0000	200/200	100,0000
A25	197/200	98,5000	197/200	98,5000
A26	198/200	99,0000	198/200	99,0000
A27	198/200	99,0000	198/200	99,0000
A28	197/200	98,5000	197/200	98,5000
A29	197/200	98,5000	196/200	98,0000
A30	198/200	99,0000	198/200	99,0000

FONTE: A autora (2016).

GRÁFICO 5 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *CIRCLE AND SQUARE*



FONTE: A autora (2016)

TABELA 37 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *IRIS*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A1	BG	1,0000	0,5000	98,3333	37	30,8333
	GQ	90,5097	0,0405	96,6667	16	13,3333
A2	BG/GQ	2,0000	0,5000	97,5000	32	26,6667
A3	BG	181,0193	0,0625	97,5000	13	10,8333
	GQ	245,1464	0,0442	97,5000	13	10,8333
A4	BG	256,0000	0,1250	97,5000	13	10,8333
	GQ	256,0000	0,1277	97,5000	13	10,8333
A5	BG	256,0000	0,0156	96,6667	17	14,1667
	GQ	152,2185	0,0274	96,6667	17	14,1667
A6	BG	32,0000	0,0078	98,3333	41	34,1667
	GQ	3,2210	0,0811	98,3333	41	34,1667
A7	BG	128,0000	0,0884	97,5000	14	11,6667
	GQ	142,6415	0,0884	97,5000	14	11,6667
A8	BG	256,0000	0,0442	96,6667	14	11,6667
	GQ	117,3765	0,0884	96,6667	14	11,6667
A9	BG	11,3137	0,7071	98,3333	21	17,5000
	GQ	107,6347	0,1250	97,5000	14	11,6667
A10	BG	2,8284	0,3536	98,3333	29	24,1667
	GQ	45,2548	0,0055	97,5000	42	35,0000
A11	BG	256,0000	0,0039	96,6667	26	21,6667
	GQ	45,2548	0,0221	96,6667	26	21,6667
A12	BG	128,0000	0,1250	97,5000	13	10,8333
	GQ	0,2500	2,8284	97,5000	57	47,5000
A13	BG/GQ	90,5097	0,0442	98,3333	17	14,1667
A14	BG/GQ	256,0000	0,1250	99,1667	13	10,8333
A15	BG	32,0000	0,0313	97,5000	25	20,8333
	GQ	22,6274	0,0442	97,5000	25	20,8333
A16	BG	22,6274	0,0442	97,5000	27	22,5000
	GQ	11,5615	0,1393	97,5000	23	19,1667
A17	BG	128,0000	0,0156	98,3333	19	15,8333
	GQ	64,0000	0,0313	98,3333	19	15,8333
A18	BG	1,0000	0,7071	95,0000	37	30,8333
	GQ	2,7085	0,1051	95,0000	42	35,0000
A19	BG	45,2548	0,0313	99,1667	20	16,6667
	GQ	24,6754	0,0653	99,1667	19	15,8333
A20	BG	256,0000	0,1768	98,3333	13	10,8333
	GQ	256,0000	0,1250	97,5000	13	10,8333
A21	BG	256,0000	0,0055	96,6667	22	18,3333
	GQ	5,6569	0,0625	96,6667	38	31,6667

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A22	BG	64,0000	0,0442	97,5000	18	15,0000
	GQ	117,3765	0,0221	97,5000	18	15,0000
A23	BG	181,0193	0,0884	97,5000	13	10,8333
	GQ	215,2695	0,0712	97,5000	13	10,8333
A24	BG	22,6274	0,0078	97,5000	47	39,1667
	GQ	128,0000	0,0884	96,6667	14	11,6667
A25	BG	0,5000	0,3536	96,6667	54	45,0000
	GQ	53,8174	0,0372	95,8333	20	16,6667
A26	BG	128,0000	0,0055	96,6667	28	23,3333
	GQ	35,6604	0,0212	96,6667	27	22,5000
A27	BG	8,0000	0,1250	98,3333	26	21,6667
	GQ	189,0338	0,0625	97,5000	13	10,8333
A28	BG	32,0000	0,0221	99,1667	27	22,5000
	GQ	16,7084	0,0442	99,1667	27	22,5000
A29	BG	45,2548	0,1768	96,6667	18	15,0000
	GQ	128,0000	0,0884	95,8333	16	13,3333
A30	BG	256,0000	0,0442	96,6667	14	11,6667
	GQ	69,7925	0,1250	96,6667	16	13,3333

FONTE: A autora (2016).

TABELA 38 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *IRIS*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	164	100	15,0597	-	84,9403
A2	1089	135	100	12,3967	-	87,6033
A3	1089	199	100	18,2737	-	81,7263
A4	1089	291	100	26,7218	-	73,2782
A5	1089	181	100	16,6208	-	83,3792
A6	1089	129	100	11,8457	-	88,1543
A7	1089	319	100	29,2929	-	70,7071
A8	1089	198	100	18,1818	-	81,8182
A9	1089	189	100	17,3554	-	82,6446
A10	1089	129	100	11,8457	-	88,1543
A11	1089	435	100	39,9449	-	60,0551
A12	1089	137	100	12,5804	-	87,4197
A13	1089	443	100	40,6795	-	59,3205
A14	1089	216	100	19,8347	-	80,1653
A15	1089	279	100	25,6198	-	74,3802
A16	1089	218	100	20,0184	-	79,9816
A17	1089	126	100	11,5703	-	88,4298
A18	1089	272	100	24,9770	-	75,0230
A19	1089	195	100	17,9063	-	82,0937
A20	1089	274	100	25,1607	-	74,8393
A21	1089	131	100	12,0294	-	87,9706
A22	1089	151	100	13,8659	-	86,1341
A23	1089	260	100	23,8751	-	76,1249
A24	1089	204	100	18,7328	-	81,2672
A25	1089	414	100	38,0165	-	61,9835
A26	1089	228	100	20,9366	-	79,0634
A27	1089	159	100	14,6006	-	85,3995
A28	1089	213	100	19,5592	-	80,4408
A29	1089	184	100	16,8962	-	83,1038
A30	1089	244	100	22,4059	-	77,5941

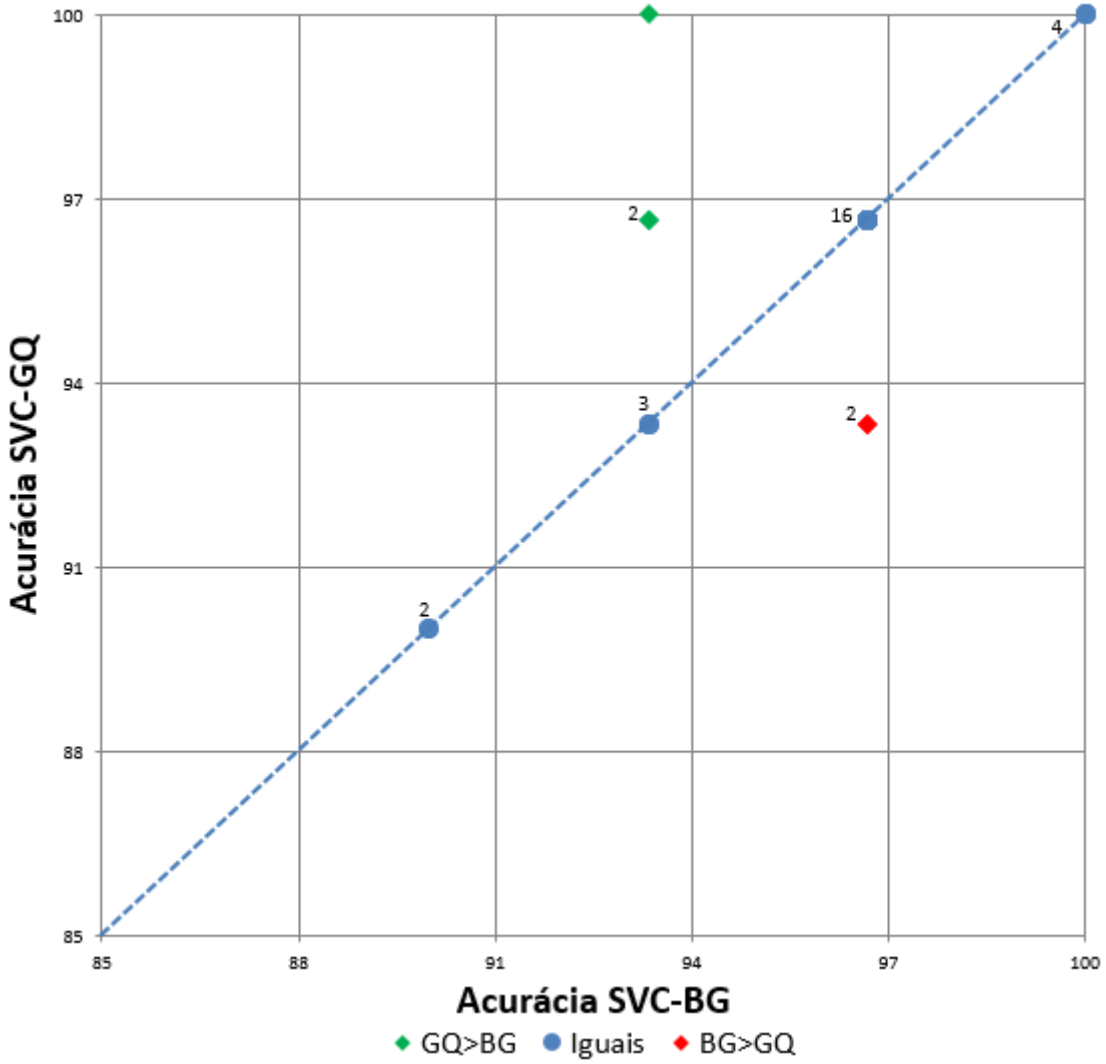
FONTE: A autora (2016).

TABELA 39 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *IRIS*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	28/30	93,3333	29/30	96,6667
A2	29/30	96,6667	29/30	96,6667
A3	28/30	93,3333	28/30	93,3333
A4	29/30	96,6667	29/30	96,6667
A5	29/30	96,6667	29/30	96,6667
A6	29/30	96,6667	29/30	96,6667
A7	29/30	96,6667	29/30	96,6667
A8	29/30	96,6667	29/30	96,6667
A9	28/30	93,3333	29/30	96,6667
A10	29/30	96,6667	29/30	96,6667
A11	29/30	96,6667	29/30	96,6667
A12	29/30	96,6667	28/30	93,3333
A13	29/30	96,6667	29/30	96,6667
A14	28/30	93,3333	28/30	93,3333
A15	29/30	96,6667	29/30	96,6667
A16	30/30	100,0000	30/30	100,0000
A17	29/30	96,6667	29/30	96,6667
A18	30/30	100,0000	30/30	100,0000
A19	27/30	90,0000	27/30	90,0000
A20	29/30	96,6667	29/30	96,6667
A21	29/30	96,6667	29/30	96,6667
A22	29/30	96,6667	29/30	96,6667
A23	28/30	93,3333	28/30	93,3333
A24	29/30	96,6667	29/30	96,6667
A25	28/30	93,3333	30/30	100,0000
A26	30/30	100,0000	30/30	100,0000
A27	29/30	96,6667	28/30	93,3333
A28	27/30	90,0000	27/30	90,0000
A29	30/30	100,0000	30/30	100,0000
A30	29/30	96,6667	29/30	96,6667

FONTE: A autora (2016).

GRÁFICO 6 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *IRIS*



FONTE: A autora (2016)

TABELA 40 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *SVMGUIDE 2*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A1	BG	256,0000	0,0156	83,7061	114	36,4217
	GQ	109,9916	0,0397	83,3866	120	38,3387
A2	BG	4,0000	0,2500	84,3450	147	46,9649
	GQ	1,2419	0,7071	84,0256	191	61,0224
A3	BG	90,5097	0,0156	82,7476	118	37,6997
	GQ	148,9568	0,0110	82,7476	116	37,0607
A4	BG	4,0000	0,2500	86,9010	144	46,0064
	GQ	20,3048	0,0682	86,2620	117	37,3802
A5	BG	2,0000	0,5000	86,5815	174	55,5911
	GQ	1,9571	0,5109	86,5815	176	56,2300
A6	BG	2,8284	0,2500	86,5815	145	46,3259
	GQ	7,1788	0,1250	85,3035	128	40,8946
A7	BG	22,6274	0,0884	84,3450	124	39,6166
	GQ	7,1788	0,2668	84,3450	148	47,2843
A8	BG	0,7071	0,5000	84,3450	180	57,5080
	GQ	5,6569	0,0625	83,7061	139	44,4090
A9	BG	11,3137	0,0625	85,3035	131	41,8530
	GQ	12,3377	0,0599	85,3035	130	41,5336
A10	BG	2,0000	0,1768	84,6645	149	47,6038
	GQ	64,0000	0,5000	84,6645	182	58,1470
A11	BG	8,0000	0,1250	85,3035	127	40,5751
	GQ	13,7490	0,1051	84,6645	121	38,6582
A12	BG	16,0000	0,0625	83,0671	126	40,2556
	GQ	256,0000	0,0153	83,0671	109	34,8243
A13	BG/GQ	16,0000	0,0625	82,4281	131	41,8530
A14	BG	2,8284	0,3536	85,9425	155	49,5208
	GQ	3,0844	0,7071	85,9425	192	61,3419
A15	BG	2,8284	0,1250	86,5815	142	45,3674
	GQ	2,7085	0,1098	86,5815	145	46,3259
A16	BG/GQ	2,0000	0,5000	86,9010	172	54,9521
A17	BG	1,4142	0,5000	84,3450	177	56,5495
	GQ	6,7272	0,0405	83,7061	142	45,3674
A18	BG	4,0000	0,0884	84,6645	143	45,6869
	GQ	245,1464	0,0078	83,7061	113	36,1022
A19	BG	22,6274	0,0442	84,6645	120	38,3387
	GQ	19,0273	0,0442	85,3035	123	39,2971
A20	BG	8,0000	0,3536	83,7061	161	51,4377
	GQ	8,8191	0,3536	84,0256	162	51,7572
A21	BG	45,2548	0,0110	85,3035	126	40,2556
	GQ	32,0000	0,0156	85,3035	127	40,5751

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A22	BG	64,0000	0,0313	86,9010	115	36,7412
	GQ	76,1093	0,0313	86,5815	112	35,7828
A23	BG/GQ	4,0000	0,2500	87,8594	140	44,7284
A24	BG	16,0000	0,0313	85,3035	131	41,8530
	GQ	16,0000	0,0442	84,9840	128	40,8946
A25	BG	181,0193	0,0156	84,3450	114	36,4217
	GQ	139,5850	0,0178	84,6645	115	36,7412
A26	BG	4,0000	0,2500	82,1086	147	46,9649
	GQ	128,0000	0,0039	81,7891	132	42,1725
A27	BG/GQ	1,4142	0,2500	84,3450	158	50,4792
A28	BG	181,0193	0,0156	85,3035	104	33,2268
	GQ	197,4030	0,0150	85,3035	103	32,9074
A29	BG	4,0000	0,3536	84,6645	165	52,7157
	GQ	1,4768	0,3692	84,6645	173	55,2716
A30	BG	2,8284	0,1250	84,3450	146	46,6454
	GQ	2,7085	0,1393	84,3450	148	47,2843

FONTE: A autora (2016).

TABELA 41 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *SVMGUIDE 2*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	160	100	14,6924	-	85,3076
A2	1089	158	100	14,5087	-	85,4913
A3	1089	273	100	25,0689	-	74,9311
A4	1089	217	100	19,9265	-	80,0735
A5	1089	162	100	14,8760	-	85,1240
A6	1089	133	100	12,2130	-	87,7870
A7	1089	195	100	17,9063	-	82,0937
A8	1089	133	100	12,2130	-	87,7870
A9	1089	153	100	14,0496	-	85,9504
A10	1089	196	100	17,9982	-	82,0018
A11	1089	140	100	12,8558	-	87,1442
A12	1089	326	100	29,9357	-	70,0643
A13	1089	125	100	11,4784	-	88,5216
A14	1089	157	100	14,4169	-	85,5831
A15	1089	197	100	18,0900	-	81,9100
A16	1089	136	100	12,4885	-	87,5115
A17	1089	221	100	20,2939	-	79,7062
A18	1089	148	100	13,5905	-	86,4096
A19	1089	155	100	14,2332	-	85,7668
A20	1089	136	100	12,4885	-	87,5115
A21	1089	130	100	11,9376	-	88,0624
A22	1089	192	100	17,6309	-	82,3692
A23	1089	135	100	12,3967	-	87,6033
A24	1089	130	100	11,9376	-	88,0624
A25	1089	241	100	22,1304	-	77,8696
A26	1089	97	100	8,9073	-	91,0927
A27	1089	137	100	12,5804	-	87,4197
A28	1089	161	100	14,7842	-	85,2158
A29	1089	178	100	16,3453	-	83,6547
A30	1089	133	100	12,2130	-	87,7870

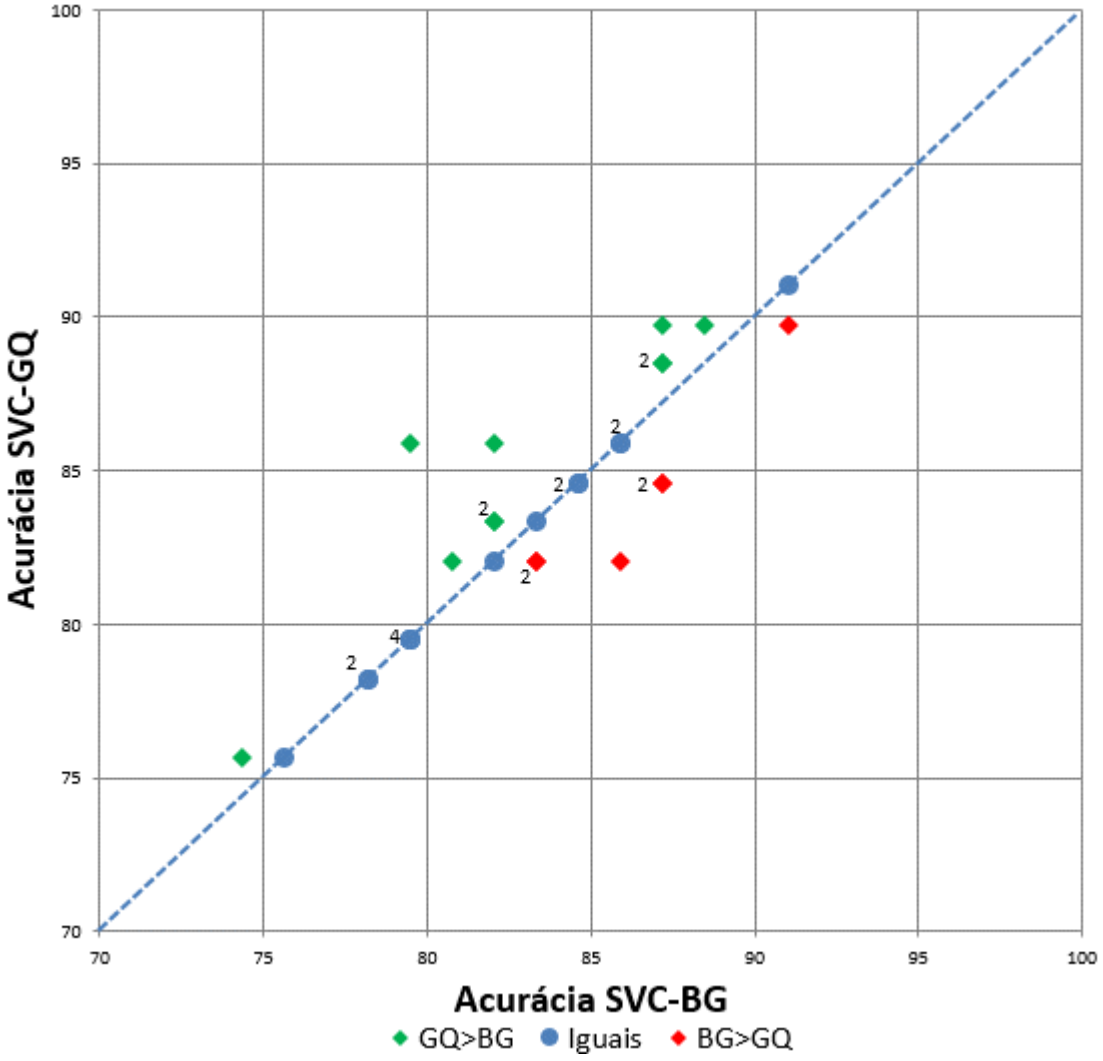
FONTE: A autora (2016).

TABELA 42 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *SVMGUIDE 2*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	69/78	88,4615	70/78	89,7436
A2	68/78	87,1795	66/78	84,6154
A3	66/78	84,6154	66/78	84,6154
A4	62/78	79,4872	62/78	79,4872
A5	67/78	85,8974	67/78	85,8974
A6	58/78	74,3590	59/78	75,6410
A7	65/78	83,3333	64/78	82,0513
A8	62/78	79,4872	67/78	85,8974
A9	67/78	85,8974	67/78	85,8974
A10	64/78	82,0513	65/78	83,3333
A11	62/78	79,4872	62/78	79,4872
A12	64/78	82,0513	65/78	83,3333
A13	71/78	91,0256	71/78	91,0256
A14	63/78	80,7692	64/78	82,0513
A15	65/78	83,3333	64/78	82,0513
A16	64/78	82,0513	64/78	82,0513
A17	68/78	87,1795	66/78	84,6154
A18	67/78	85,8974	64/78	82,0513
A19	62/78	79,4872	62/78	79,4872
A20	66/78	84,6154	66/78	84,6154
A21	59/78	75,6410	59/78	75,6410
A22	61/78	78,2051	61/78	78,2051
A23	61/78	78,2051	61/78	78,2051
A24	64/78	82,0513	67/78	85,8974
A25	71/78	91,0256	70/78	89,7436
A26	68/78	87,1795	69/78	88,4615
A27	65/78	83,3333	65/78	83,3333
A28	62/78	79,4872	62/78	79,4872
A29	68/78	87,1795	70/78	89,7436
A30	68/78	87,1795	69/78	88,4615

FONTE: A autora (2016).

GRÁFICO 7 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A SVMGUIDE 2



FONTE: A autora (2016)

TABELA 43 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *VEHICLE*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A1	BG	256,0000	0,0625	87,2969	247	36,4845
	GQ	181,0193	0,0884	87,2969	256	37,8139
A2	BG	181,0193	0,1768	84,7858	252	37,2230
	GQ	96,5865	0,0884	84,7858	263	38,8479
A3	BG	90,5097	0,1250	86,7061	263	38,8479
	GQ	197,4030	0,0811	86,4106	250	36,9276
A4	BG/GQ	256,0000	0,0884	84,6381	246	36,3368
A5	BG	128,0000	0,1250	83,7518	262	38,7001
	GQ	79,4789	0,1146	83,3087	271	40,0295
A6	BG	181,0193	0,0884	84,7858	240	35,4505
	GQ	206,1428	0,1768	84,7858	243	35,8937
A7	BG	256,0000	0,2500	84,3427	262	38,7001
	GQ	256,0000	0,2013	84,3427	257	37,9616
A8	BG/GQ	256,0000	0,0884	83,8996	249	36,7799
A9	BG	256,0000	0,0884	85,2290	243	35,8936
	GQ	256,0000	0,0526	85,3767	248	36,6322
A10	BG	256,0000	0,0625	83,3087	253	37,3708
	GQ	152,2185	0,0811	83,7518	261	38,5524
A11	BG	128,0000	0,0884	85,0812	255	37,6662
	GQ	206,1428	0,0682	84,7858	248	36,6322
A12	BG	256,0000	0,0884	85,2290	243	35,8936
	GQ	206,1428	0,0884	85,3767	249	36,7799
A13	BG/GQ	256,0000	0,0313	83,7518	263	38,8479
A14	BG	256,0000	0,0884	84,7858	244	36,0414
	GQ	197,4030	0,1098	84,7858	248	36,6322
A15	BG	256,0000	0,1768	84,4904	246	36,3368
	GQ	181,0193	0,2013	84,3427	253	37,3708
A16	BG	128,0000	0,2500	85,6721	271	40,0295
	GQ	234,7530	0,1768	85,6721	254	37,5185
A17	BG	181,0193	0,0884	85,3767	252	37,2230
	GQ	96,5865	0,1250	85,2290	268	39,5864
A18	BG	90,5097	0,2500	84,6381	268	39,5864
	GQ	66,8335	0,2668	84,7858	276	40,7681
A19	BG	181,0193	0,1250	85,8198	254	37,5185
	GQ	234,7530	0,1051	86,1152	247	36,4845
A20	BG	64,0000	0,1250	85,2290	279	41,2112
	GQ	215,2695	0,2500	84,6381	263	38,8479
A21	BG/GQ	256,0000	0,0884	85,3767	245	36,1891
A22	BG	256,0000	0,1768	84,7858	250	36,9276
	GQ	128,0000	0,2243	84,6381	270	39,8818

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A23	BG	128,0000	0,0442	84,6381	271	40,0295
	GQ	184,9831	0,1305	84,7858	243	35,8937
A24	BG	90,5097	0,1768	84,0473	267	39,4387
	GQ	215,2695	0,0811	84,0473	248	36,6322
A25	BG	256,0000	0,1250	84,7858	249	36,7799
	GQ	139,5850	0,1768	84,3427	261	38,5524
A26	BG	181,0193	0,0442	83,6041	274	40,4727
	GQ	184,9831	0,0442	83,6041	273	40,3250
A27	BG	128,0000	0,1250	85,2290	265	39,1433
	GQ	206,1428	0,1250	84,9335	258	38,1093
A28	BG	32,0000	0,1250	85,0812	294	43,4269
	GQ	36,4412	0,1250	84,4904	288	42,5406
A29	BG	256,0000	0,0625	85,3767	245	36,1891
	GQ	224,8003	0,0760	85,2290	243	35,8937
A30	BG	90,5097	0,1250	85,6721	259	38,2570
	GQ	66,8335	0,1487	85,6721	266	39,2910

FONTE: A autora (2016).

TABELA 44 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *VEHICLE*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	167	100	15,3352	-	84,6648
A2	1089	131	100	12,0294	-	87,9706
A3	1089	215	100	19,7429	-	80,2571
A4	1089	217	100	19,9265	-	80,0735
A5	1089	237	100	21,7631	-	78,2369
A6	1089	198	100	18,1818	-	81,8182
A7	1089	187	100	17,1717	-	82,8283
A8	1089	200	100	18,3655	-	81,6345
A9	1089	310	100	28,4665	-	71,5335
A10	1089	192	100	17,6309	-	82,3691
A11	1089	200	100	18,3655	-	81,6345
A12	1089	222	100	20,3857	-	79,6143
A13	1089	217	100	19,9265	-	80,0735
A14	1089	205	100	18,8246	-	81,1754
A15	1089	187	100	17,1717	-	82,8283
A16	1089	204	100	18,7328	-	81,2672
A17	1089	131	100	12,0294	-	87,9706
A18	1089	195	100	17,9063	-	82,0937
A19	1089	210	100	19,2838	-	80,7163
A20	1089	214	100	19,6511	-	80,3489
A21	1089	219	100	20,1102	-	79,8898
A22	1089	178	100	16,3453	-	83,6547
A23	1089	275	100	25,2525	-	74,7475
A24	1089	198	100	18,1818	-	81,8182
A25	1089	203	100	18,6410	-	81,3590
A26	1089	188	100	17,2635	-	82,7365
A27	1089	186	100	17,0799	-	82,9201
A28	1089	303	100	27,8237	-	72,1763
A29	1089	224	100	20,5693	-	79,4307
A30	1089	225	100	20,6612	-	79,3388

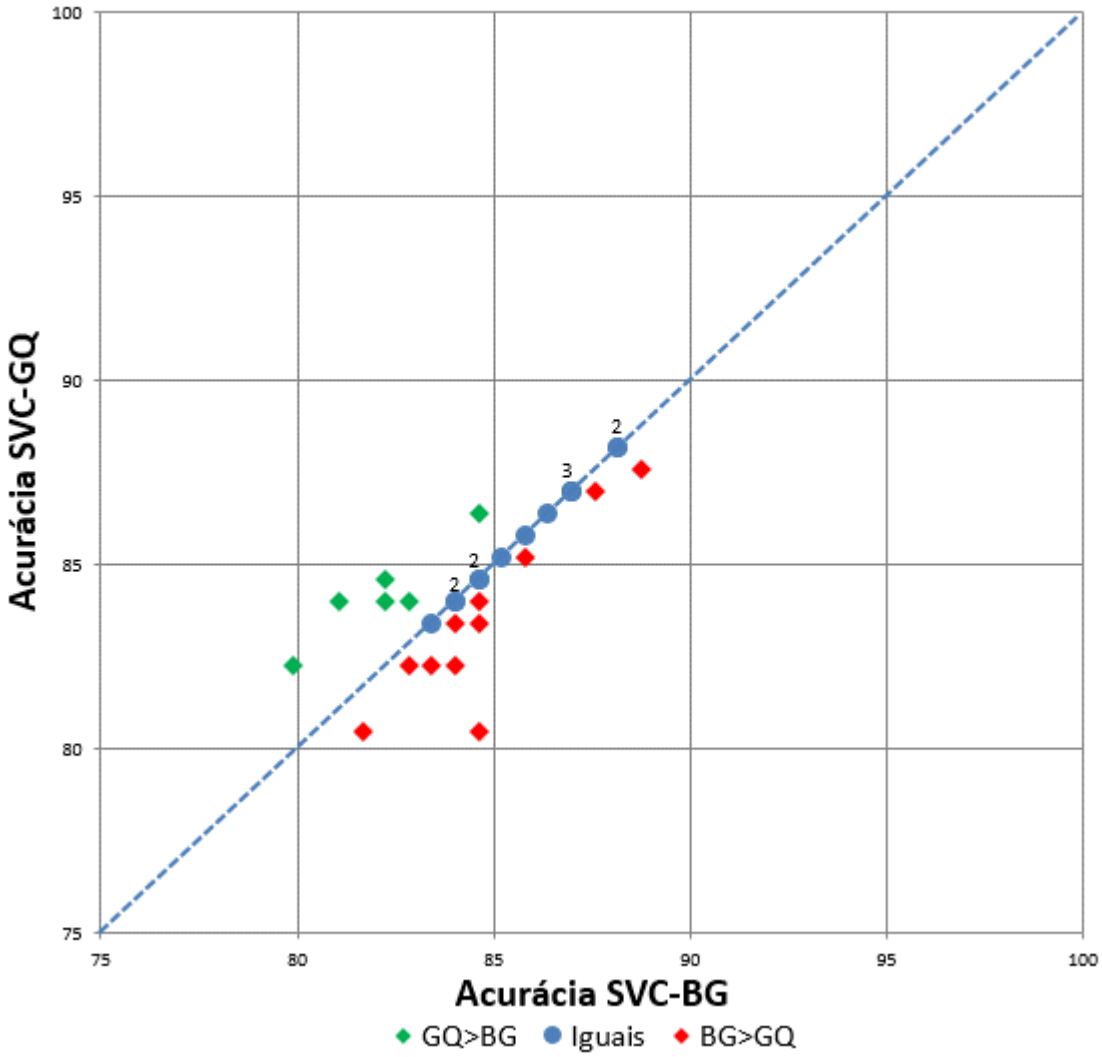
FONTE: A autora (2016).

TABELA 45 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *VEHICLE*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	139/169	82,2485	142/169	84,0237
A2	139/169	82,2485	143/169	84,6154
A3	140/169	82,8402	142/169	84,0237
A4	147/169	86,9822	147/169	86,9822
A5	149/169	88,1657	149/169	88,1657
A6	143/169	84,6154	143/169	84,6154
A7	141/169	83,4320	141/169	83,4320
A8	149/169	88,1657	149/169	88,1657
A9	138/169	81,6568	136/169	80,4734
A10	150/169	88,7574	148/169	87,5740
A11	143/169	84,6154	141/169	83,4320
A12	143/169	84,6154	143/169	84,6154
A13	144/169	85,2071	144/169	85,2071
A14	142/169	84,0237	142/169	84,0237
A15	148/169	87,5740	147/169	86,9822
A16	142/169	84,0237	141/169	83,4320
A17	143/169	84,6154	142/169	84,0237
A18	142/169	84,0237	139/169	82,2485
A19	145/169	85,7988	145/169	85,7988
A20	143/169	84,6154	136/169	80,4734
A21	147/169	86,9822	147/169	86,9822
A22	143/169	84,6154	146/169	86,3905
A23	135/169	79,8817	139/169	82,2485
A24	137/169	81,0651	142/169	84,0237
A25	140/169	82,8402	139/169	82,2485
A26	147/169	86,9822	147/169	86,9822
A27	146/169	86,3905	146/169	86,3905
A28	145/169	85,7988	144/169	85,2071
A29	142/169	84,0237	142/169	84,0237
A30	141/169	83,4320	139/169	82,2485

FONTE: A autora (2016).

GRÁFICO 8 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *VEHICLE*



FONTE: A autora (2016).

TABELA 46 – PARÂMETROS ENCONTRADOS PELA BG E GQ PARA O CONJUNTO *SEGMENT*

continua

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A1	BG	64,0000	1,0000	97,2403	345	18,6688
	GQ	36,4412	0,1250	97,0779	289	15,6385
A2	BG	256,0000	0,1250	97,2403	239	12,9329
	GQ	117,3765	0,1768	97,1861	257	13,9069
A3	BG/GQ	90,5097	0,3536	97,5108	274	14,8268
A4	BG	11,3137	1,0000	97,1861	386	20,8874
	GQ	162,4385	0,7384	97,1320	326	17,6407
A5	BG	64,0000	0,1768	97,1320	280	15,1515
	GQ	45,2548	0,2500	96,9697	294	15,9091
A6	BG	22,6274	0,5000	97,2944	328	17,7489
	GQ	48,2933	0,3536	97,2403	297	16,0714
A7	BG	128,0000	0,3536	97,3485	274	14,8268
	GQ	32,0000	1,0000	97,1320	365	19,7511
A8	BG	11,3137	1,0000	97,5108	380	20,5628
	GQ	197,4030	0,2500	97,2944	255	13,7987
A9	BG	256,0000	0,2500	97,2944	261	14,1234
	GQ	90,5097	0,3692	97,1861	288	15,5844
A10	BG	32,0000	0,5000	97,1320	326	17,6407
	GQ	35,6604	0,4685	97,1861	322	17,4242
A11	BG	22,6274	1,0000	97,8355	368	19,9134
	GQ	215,2695	0,5000	97,8355	285	15,4221
A12	BG	128,0000	0,5000	97,1320	308	16,6667
	GQ	224,8003	0,3242	97,1320	276	14,9351
A13	BG	256,0000	1,0000	97,2403	352	19,0476
	GQ	50,4314	0,5000	97,1320	317	17,1537
A14	BG	256,0000	0,2500	97,0779	262	14,1775
	GQ	148,9568	0,3242	97,0238	275	14,8810
A15	BG	256,0000	0,3536	96,9156	272	14,7186
	GQ	117,3765	0,7071	96,8074	322	17,4242
A16	BG	90,5097	1,0000	97,2944	350	18,9394
	GQ	98,7015	0,5000	97,1861	297	16,0714
A17	BG	22,6274	1,0000	97,5108	363	19,6429
	GQ	56,2001	0,7071	97,4026	316	17,0996
A18	BG	256,0000	0,1768	97,2403	242	13,0952
	GQ	152,2185	0,2500	97,2944	254	13,7446
A19	BG	64,0000	0,5000	97,4567	308	16,6667
	GQ	152,2185	0,3692	97,4567	278	15,0433
A20	BG	90,5097	0,7071	97,2403	333	18,0195
	GQ	173,3447	0,0811	97,1861	265	14,3398

conclusão

CONJUNTO	MÉTODO	C	γ	TAXA VC (%)	VS	VS (%)
A21	BG	90,5097	0,1250	97,1861	270	14,6104
	GQ	152,2185	0,0884	97,0779	261	14,1234
A22	BG	16,0000	0,7071	97,1320	356	19,2641
	GQ	128,0000	0,2500	97,0238	265	14,3398
A23	BG	128,0000	0,2500	97,1861	274	14,8268
	GQ	206,1428	0,2500	97,1861	267	14,4481
A24	BG	32,0000	0,7071	97,2944	335	18,1277
	GQ	34,8963	0,7071	97,2944	334	18,0736
A25	BG/GQ	90,5097	0,5000	97,1320	297	16,0714
A26	BG	181,0193	1,4142	96,9697	390	21,1039
	GQ	256,0000	0,1487	96,8615	249	13,4740
A27	BG	11,3137	1,0000	97,6190	381	20,6169
	GQ	22,6274	0,7071	97,2944	344	18,6147
A28	BG	64,0000	0,5000	97,5649	301	16,2879
	GQ	64,0000	0,5109	97,5649	302	16,3420
A29	BG	128,0000	0,3536	96,9156	280	15,1515
	GQ	133,6670	0,3536	96,9156	279	15,0974
A30	BG	256,0000	0,3536	96,8074	276	14,9351
	GQ	165,9955	0,5000	96,9156	301	16,2879

FONTE: A autora (2016).

TABELA 47 – COMPARAÇÃO ENTRE O NÚMERO DE OPERAÇÕES REALIZADAS PELOS MÉTODOS BG E GQ PARA O CONJUNTO *SEGMENT*

CONJUNTO	Nº OPERAÇÕES		OPERAÇÕES (%)		REDUÇÃO DE OPERAÇÕES (%)	
	BG	GQ	BG	GQ	BG	GQ
A1	1089	130	100	11,9376	-	88,0624
A2	1089	154	100	14,1414	-	85,8586
A3	1089	133	100	12,2130	-	87,7870
A4	1089	187	100	17,1717	-	82,8283
A5	1089	133	100	12,2130	-	87,7870
A6	1089	140	100	12,8558	-	87,1442
A7	1089	143	100	13,1313	-	86,8687
A8	1089	180	100	16,5289	-	83,4711
A9	1089	141	100	12,9477	-	87,0523
A10	1089	206	100	18,9164	-	81,0836
A11	1089	213	100	19,5592	-	80,4408
A12	1089	363	100	33,3333	-	66,6667
A13	1089	138	100	12,6722	-	87,3278
A14	1089	209	100	19,1919	-	80,8081
A15	1089	177	100	16,2534	-	83,7466
A16	1089	162	100	14,8760	-	85,1240
A17	1089	142	100	13,0395	-	86,9605
A18	1089	184	100	16,8962	-	83,1038
A19	1089	169	100	15,5188	-	84,4812
A20	1089	219	100	20,1102	-	79,8898
A21	1089	155	100	14,2332	-	85,7668
A22	1089	130	100	11,9376	-	88,0624
A23	1089	249	100	22,8650	-	77,1350
A24	1089	164	100	15,0597	-	84,9403
A25	1089	170	100	15,6107	-	84,3894
A26	1089	206	100	18,9164	-	81,0836
A27	1089	133	100	12,2130	-	87,7870
A28	1089	130	100	11,9376	-	88,0624
A29	1089	133	100	12,2130	-	87,7870
A30	1089	178	100	16,3453	-	83,6547

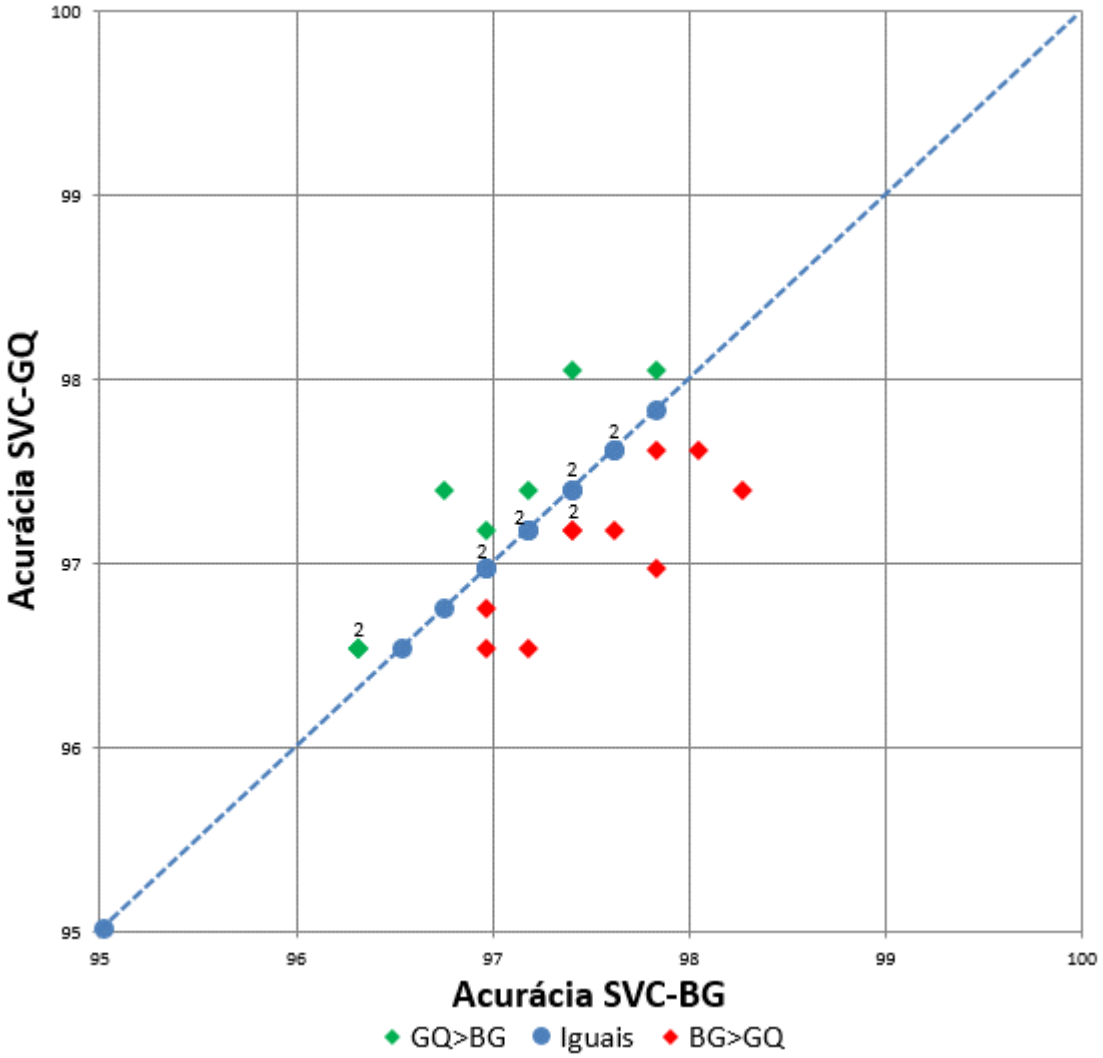
FONTE: A autora (2016).

TABELA 48 – QA E ACURÁCIA DO SVC PARA O CONJUNTO *SEGMENT*

CONJUNTO	BUSCA POR *GRID*		*GRID-QUADTREE*	
	QA	ACURÁCIA (%)	QA	ACURÁCIA (%)
A1	445/462	96,3203	446/462	96,5368
A2	448/462	96,9697	448/462	96,9697
A3	439/462	95,0216	439/462	95,0216
A4	452/462	97,8355	451/462	97,6190
A5	447/462	96,7532	450/462	97,4026
A6	450/462	97,4026	450/462	97,4026
A7	450/462	97,4026	449/462	97,1861
A8	449/462	97,1861	446/462	96,5368
A9	451/462	97,6190	451/462	97,6190
A10	451/462	97,6190	451/462	97,6190
A11	449/462	97,1861	449/462	97,1861
A12	448/462	96,9697	446/462	96,5368
A13	454/462	98,2684	450/462	97,4026
A14	452/462	97,8355	452/462	97,8355
A15	451/462	97,6190	449/462	97,1861
A16	450/462	97,4026	449/462	97,1861
A17	446/462	96,5368	446/462	96,5368
A18	445/462	96,3203	446/462	96,5368
A19	449/462	97,1861	450/462	97,4026
A20	452/462	97,8355	448/462	96,9697
A21	450/462	97,4026	450/462	97,4026
A22	447/462	96,7532	447/462	96,7532
A23	450/462	97,4026	453/462	98,0519
A24	448/462	96,9697	448/462	96,9697
A25	449/462	97,1861	449/462	97,1861
A26	452/462	97,8355	453/462	98,0519
A27	445/462	96,3203	446/462	96,5368
A28	448/462	96,9697	447/462	96,7532
A29	448/462	96,9697	449/462	97,1861
A30	453/462	98,0519	451/462	97,6190

FONTE: A autora (2016).

GRÁFICO 9 - COMPARAÇÃO ENTRE AS ACURÁCIAS DO SVC-BG E SVC-GQ PARA A *SEGMENT*



FONTE: A autora (2016).

QUADRO 28 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS EMPREGADAS NO ESTUDO ESTATÍSTICO DE P

REFERÊNCIAS	SONAR	AUSTRALIAN	FOURCLASS	GERMAN	WINE	GLASS
Ho e Kleinberg (1996)			X			
Hsu e Lin (2002)					X	X
Huang e Dun (2008)				X		
Huang e Wang (2006)	X	X		X		
Kapp, Sabourin e Maupin (2012)				X		
Lessmann, Stahlbock e Crone (2006)		X		X		
Lin *et al.* (2008a)	X	X		X		X
Lin *et al.* (2008b)	X	X		X	X	X
Lin *et al.* (2015)		X		X	X	X
Liu e Xu (2013)	X				X	X
Miranda *et al*. (2014)	X				X	X
Pang *et al.* (2011)				X		
Varewyck e Martens (2011)		X				
Wang, Huang e Cheng (2014)					X	
Zhao *et al*. (2011)	X	X		X	X	

Fonte: A autora (2016).

QUADRO 29 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS DO GRUPO 1

Continua

REFERÊNCIAS	LIVER-LISORDERS	IONOSPHERE	BREAST CANCER	DIABETES	CIRCLE AND SQUARE	MUSHROOM	IRIS	SVMGUIDE 2	VEHICLE	SEGMENT
Carpenter, Grossberg e Reynolds (1991)					X					
Chapelle *et al.* (2002)			X	X						
Guo, Yang e Xiao (2010)							X			
Hsu e Lin (2002)							X		X	X
Hsu, Chang e Lin (2010)								X		
Huang e Wang (2006)		X	X				X		X	
Jiang e Zou (2013)				X						
Kapp, Sabourin e Maupin (2012)					X	X				X
Keerthi e Lin (2003)				X						
Lessmann, Stahlbock e Crone (2006)			X							
Li *et al.* (2012)		X					X			
Lin *et al.* (2008a)		X	X						X	
Lin *et al.* (2008b)		X	X				X		X	
Lin *et al.* (2015)		X							X	
Liu e Xu (2013)		X					X			
Miranda *et al*. (2014)		X		X		X	X		X	X
Pang *et al.* (2011)				X						

Conclusão

Varewyck e Martens (2011)				X		X				X
Wang, Huang e Cheng (2014)							X			X
Zhang, Chen e He (2010)			X	X						
Zhao *et al*. (2011)	X	X	X	X			X		X	
Zhao *et al*. (2012)				X						

Fonte: A autora (2016).

QUADRO 30 – REFERÊNCIAS QUE AVALIARAM AS BASES DE DADOS DO GRUPO 2

REFERÊNCIAS	A1A	SPLICE	SVMGUIDE 1	SVMGUIDE 3	W1A	DNA	SATIMAGE	SVMGUIDE 4	PENDIGITS	VOWEL
Hsu e Lin (2002)						X	X			X
Hsu, Chang e Lin (2010)			X	X				X		
Huang e Wang (2006)										X
Kapp, Sabourin e Maupin (2012)	X	X	X			X	X			
Keerthi e Lin (2003)	X	X								
Lebrun *et al.* (2008)	X	X								
Lin *et al.* (2008a)										X
Lin *et al.* (2008b)										X
Lin *et al.* (2015)										X
Miranda *et al*. (2014)		X							X	
Pang *et al.* (2011)	X		X							
Platt (1998)					X					
Varewyck e Martens (2011)		X				X	X			
Wang, Huang e Cheng (2014)							X			X
Zhao *et al*. (2011)		X								X
Zhao *et al*. (2012)		X								

Fonte: A autora (2016).