

MECANISMO IMUNO-SUPRESSOR PARA SELEÇÃO DE DADOS DE  
TREINAMENTO EM PROBLEMAS DE CLASSIFICAÇÃO

Grazziela Patrocinio Figueredo

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS  
PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE  
FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS  
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM  
ENGENHARIA CIVIL.

Aprovada por:

---

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

---

Prof. Helio José Correa Barbosa, D.Sc.

---

Prof. Beatriz de Souza Leite Pires de Lima, D.Sc.

---

Prof. Luís Alfredo Vidal de Carvalho, D.Sc.

---

Prof. Marley Maria Bernardes Rebuzzi Vellasco, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO DE 2008

# **Livros Grátis**

<http://www.livrosgratis.com.br>

Milhares de livros grátis para download.

FIGUEREDO, GRAZZIELA PATROCINIO

Mecanismo Imuno-supressor para Seleção de  
Dados de Treinamento em Problemas de Classifi-  
cação [Rio de Janeiro] 2008

XV, 101 p. 29,7 cm (COPPE/UFRJ, D.Sc., En-  
genharia Civil, 2008)

Tese - Universidade Federal do Rio de  
Janeiro, COPPE

1. Sistema Imunológico
2. Sistemas Imunológicos Artificiais
3. Classificação de Dados
4. Descoberta de Conhecimento

I. COPPE/UFRJ II. Título ( série )

*Ao meu avô, José do Patrocínio.*

Na desesperação já repousava  
O peito longamente magoado,  
E, com seu dano eterno concentrado,  
Já não temia, já não desejava;

Quando ùa sombra vã me assegurava  
Que algum bem me podia estar guardado  
Em tão fermosa imagem, que o treslado  
Na alma ficou, que nela se enlevava.

;Que crédito que dá tão facilmente  
O coração àquilo que deseja,  
Quando lhe esquece o fero seu destino!

Ah! deixem-me enganar, que eu sou contente;  
Que, posto que maior dano seja,  
Fica-me a glória do que imagino.

*Camões*

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

## MECANISMO IMUNO-SUPRESSOR PARA SELEÇÃO DE DADOS DE TREINAMENTO EM PROBLEMAS DE CLASSIFICAÇÃO

Grazziela Patrocínio Figueredo

Setembro/2008

Orientadores: Nelson Francisco Favilla Ebecken e Helio José Correa Barbosa

Programa: Engenharia Civil

Os mecanismos de auto-regulação do sistema imunológico cuidam para que o excesso de células de defesa sejam eliminadas. Estes mecanismos são responsáveis, também, pela eliminação de células auto-reativas ou de baixa aptidão para agir como frente de defesa.

Partindo do estudo dos mecanismos envolvidos no processo de supressão e auto-regulação dos sistemas imunológicos, são propostos, neste trabalho, dois algoritmos imunológicos. Um com o propósito de classificação e outro para seleção de dados de treinamento.

O classificador proposto trabalha com a seleção de um subconjunto de treino significativo de forma que a construção do modelo final preditor é baseada apenas neste subconjunto. Como um aprimoramento desta técnica, concebeu-se um mecanismo para seleção de dados de treinamento.

A seleção dos dados é feita por um mecanismo inspirado no comportamento dos linfonodos e células T com relação a anticorpos pouco adaptados, que são responsáveis por gerar estímulos de sobrevivência para os plasmócitos. O objetivo foi construir um mecanismo para ser hibridizado a métodos de classificação de modo que se aumente o desempenho na construção do modelo preditor, sem que isto acarretasse em perdas na sua acurácia. Os resultados apresentados mostram que o mecanismo supressor é eficiente e adequado a diversos tipos de classificadores.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

## IMMUNO SUPPRESSIVE MECHANISM TO SELECT TRAINING DATA IN CLASSIFICATION PROBLEMS

Grazziela Patrocínio Figueredo

September/2008

Advisors: Nelson Francisco Favilla Ebecken and Helio José Correa Barbosa

Department: Civil Engineering

The immune system's self regulation mechanisms help to eliminate excessive defense cells in the organism. They are also responsible for the elimination of self-reactive cells or not well adapted for defense ones.

From the study of the mechanisms involved in the process of suppression and self-regulation of the immune systems, it is proposed, in this work, two immune based algorithms. The first one has the purpose of classification. The second one has the purpose of selecting training data.

The proposed classifier works with the selection of a training subset for the construction of the final predictor. As an improvement of this technique, it was created a selection mechanism for training data.

Data selection is made by a technique inspired by the behavior of lymphnodes and T cells towards not well adapted antibodies. They are responsible for emitting survival signals for plasmocytes. The objective was to build a selector mechanism to work together with classifying methods in order to improve the performance on the model's construction, without loss of accuracy. Results presented show that the suppressor mechanism is effective and adequate to several types of classifiers.

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
<b>2</b>	<b>Conceitos Básicos sobre Imunologia</b>	<b>3</b>
2.1	Introdução . . . . .	3
2.2	Aspectos Gerais das Respostas Imunitárias . . . . .	4
2.3	Imunidade Natural e Adquirida . . . . .	5
2.4	Resposta Imunitária Humoral . . . . .	5
2.5	Resposta Imunitária Mediada por Células . . . . .	8
2.6	Características Gerais das Respostas Imunes . . . . .	9
2.7	Mecanismo das Respostas Imunitárias . . . . .	9
2.8	Antígeno e Antigenicidade . . . . .	11
2.9	Células e Tecidos do Sistema Imune . . . . .	12
2.9.1	O Tecido Linfóide . . . . .	13
2.9.2	Os Órgãos Linfóides . . . . .	13
2.9.3	Órgãos Linfóides Primários . . . . .	15
2.9.4	Órgãos Linfóides Secundários . . . . .	16
2.9.5	Os Linfócitos . . . . .	17
2.10	O Anticorpo . . . . .	22
2.10.1	A Interação da Molécula do Anticorpo com o Antígeno Específico	23
2.10.2	Região Variável . . . . .	24
2.10.3	Região de Dobradiça . . . . .	25
2.10.4	Regiões Constantes . . . . .	25
2.11	A Geração da Diversidade dos Anticorpos . . . . .	25
2.11.1	A Diversidade na Região V . . . . .	26
2.11.2	Processos Geradores da Diversidade dos Anticorpos . . . . .	29
2.11.3	Hipermutações Somáticas . . . . .	29
2.12	A Geração de Diversidade dos Receptores das Células T (TCR) . . . . .	30
2.13	Distinção Próprio/Não Próprio . . . . .	31
2.13.1	Seleção Positiva e Seleção Negativa . . . . .	32
2.14	A Teoria da Seleção Clonal . . . . .	32

2.15	A Arquitetura do SI e o Sistema Global de Defesa . . . . .	33
<b>3</b>	<b>Sistemas Imunológicos Artificiais</b>	<b>35</b>
3.1	Introdução . . . . .	35
3.2	Modelos do Sistema Imunológico . . . . .	36
3.2.1	A Teoria das Redes Idiotípicas . . . . .	36
3.2.2	O Modelo <i>Shape Space</i> . . . . .	36
3.2.3	Modelos Evolucionários . . . . .	37
3.3	Algoritmos e Processos . . . . .	38
3.3.1	Algoritmo CLONALG . . . . .	38
3.3.2	Algoritmo <i>OPT-AINET</i> . . . . .	38
3.4	SI e Complexidade . . . . .	40
<b>4</b>	<b>Classificação</b>	<b>42</b>
4.1	Introdução . . . . .	42
4.2	O Problema do Desbalanceamento de Classes . . . . .	43
4.2.1	Métricas de Avaliação de Classificadores . . . . .	43
4.2.2	A Técnica <i>Bagging</i> . . . . .	46
4.2.3	<i>SMOTE Synthetic Minority Over-sampling Technique</i> . . . . .	48
4.2.4	O Problema de Exemplos Irrelevantes . . . . .	48
<b>5</b>	<b>O Algoritmo SUPRAIC (<i>Suppressor Artificial Immune Classifier</i>)</b>	<b>51</b>
5.1	Introdução . . . . .	51
5.1.1	O Comportamento dos Sistemas Imunitários Biológicos . . . . .	52
5.2	O Algoritmo Proposto . . . . .	52
5.3	Experimentos e Resultados . . . . .	54
5.3.1	Base de Dados Câncer . . . . .	54
5.3.2	Base de Dados Diabetes dos Índios Pima . . . . .	57
5.3.3	Resultados para Outras Bases de Dados Estudadas . . . . .	58
5.4	Considerações a Respeito do Modelo . . . . .	61
<b>6</b>	<b>Mecanismo Seletor de Dados Baseado em Imuno-supressão</b>	<b>63</b>
6.1	Introdução . . . . .	63
6.2	O Mecanismo Supressor para a Seleção de Dados de Treinamento ( <i>SeleSup</i> )	64
6.3	O Algoritmo . . . . .	65
6.4	Resultados com <i>Bagging</i> . . . . .	67
6.5	Resultados com <i>Bagging</i> e SMOTE . . . . .	76
6.6	Considerações Finais a Respeito do Mecanismo Seletor de Dados . . . . .	77
<b>7</b>	<b>Considerações Finais</b>	<b>82</b>

<b>A Bases de Dados Utilizadas</b>	<b>84</b>
<b>Referências Bibliográficas</b>	<b>94</b>

# Lista de Figuras

2.1	Cinética da resposta imunitária observando as concentrações de anticorpo no soro (adaptado de [77]). . . . .	8
2.2	Aspectos essenciais das respostas imunitárias (adaptado de [77]). . . . .	10
2.3	Hematopoiese (adaptado de [38]). . . . .	12
2.4	Células do SI inato e suas funções (adaptado de [38]). . . . .	14
2.5	Distribuição dos tecidos linfóides (adaptado de [38]). . . . .	15
2.6	Esquema de um linfonodo (adaptado de [38]). . . . .	16
2.7	O baço (adaptado de [38]). . . . .	17
2.8	Formação do líquido tecidual. (adaptado de [10]). . . . .	18
2.9	O caminho percorrido pelos linfócitos e seu encontro com os antígenos. (adaptado de [38]). . . . .	18
2.10	Esquema de um TCR (adaptado de [38]). . . . .	20
2.11	Semelhança entre os TCR e as regiões Fab das Imunoglobulinas (adaptado de [38]). . . . .	21
2.12	O anticorpo considerando sua divisão em regiões (esquerda) e cadeias (direita) (adaptado de [38]). Cada braço do Y é formado pela associação de uma cadeia leve com a metade amino-terminal de uma cadeia pesada, enquanto a perna do Y é formada pelo pareamento das metades carboxi-terminais das duas cadeias pesadas [38]. . . . .	23
2.13	Organização estrutural das principais classes das Igs (adaptado de [38]). . . . .	23
2.14	O anticorpo (adaptado de [77]). . . . .	24
2.15	A geração da diversidade das regiões V do anticorpo (adaptado de [38]). . . . .	27
2.16	Organização dos segmentos gênicos de cadeias leve e pesada dentro dos grupos $\lambda$ (cromossomo 22), $\kappa$ (cromossomo 2) e H (cromossomo 14) em humanos (adaptado de [38]). . . . .	28
2.17	Disposição dos segmentos gênicos de cadeias leves e pesadas em humanos (adaptado de [38]). . . . .	28
2.18	Geração da diversidade dos TCR (adaptado de [38]). . . . .	31

3.1	Esquema de um <i>Shape space</i> , onde ● representam os anticorpos e × os antígenos. Os círculos que envolvem os anticorpos são a sua área de atuação. Qualquer antígeno presente em uma dessas áreas terá um anticorpo correspondente que irá neutralizá-lo. Quanto mais próximo o antígeno estiver do anticorpo, maior será a força de ligação. . . . .	37
4.1	Exemplos de curvas ROC para classificadores A, B, C, D e E. . . . .	45
4.2	Exemplos de AUCs para dois classificadores. . . . .	45
5.1	O modelo <i>Shape Space</i> . . . . .	53
5.2	Diagrama de Voronoi: uma representação do processo de classificação. . .	53
5.3	Quantidade total de anticorpos (dados de treinamento) para a base de dados Câncer. . . . .	56
5.4	Anticorpos da base de dados Câncer após supressão e divisão de porcentagens em 60–40%. . . . .	56
5.5	A quantidade total de anticorpos (dados de treinamento) para a base de dados Diabetes. . . . .	57
5.6	Anticorpos da base de dados Diabetes depois da supressão com divisão de porcentagens em 60–40%. . . . .	58
6.1	Esquema do funcionamento do mecanismo de seleção. . . . .	65
A.1	Distribuição das classes da base de dados Câncer. . . . .	85
A.2	Distribuição das classes da base de dados Diabetes. . . . .	86
A.3	Distribuição das classes da base de dados <i>Monks-2</i> . . . . .	86
A.4	Distribuição das classes da base de dados <i>Monks-3</i> . . . . .	87
A.5	Distribuição das classes da base de dados <i>Ionosphere</i> . . . . .	88
A.6	Distribuição das classes da base de dados Íris. . . . .	88
A.7	Distribuição das classes da base de dados <i>Well</i> . . . . .	89
A.8	Distribuição das classes da base de dados Contraceptivos. . . . .	90
A.9	Distribuição das classes da base de dados <i>Wine</i> . . . . .	90
A.10	Distribuição das classes da base de dados <i>Glass</i> . . . . .	91
A.11	Distribuição das classes da base de dados <i>Yeast</i> . . . . .	92

# Lista de Tabelas

5.1	Resultados do classificador sem utilizar o mecanismo de supressão para a base de dados Câncer com 419 anticorpos. . . . .	55
5.2	Resultados do classificador utilizando o mecanismo de supressão para a base de dados Câncer com 56 anticorpos. . . . .	55
5.3	Resultados do preditor sem mecanismo de supressão para base de dados Diabetes com 460 anticorpos. . . . .	57
5.4	Resultados do preditor com mecanismo de supressão para base de dados Diabetes com 80 anticorpos. . . . .	57
5.5	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados <i>Monks-2</i> . . . . .	59
5.6	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados <i>Monks-3</i> . . . . .	59
5.7	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados <i>Ionosphere</i> . . . . .	60
5.8	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados <i>Well</i> . . . . .	60
5.9	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados Contraceptivos. . . . .	60
5.10	Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados <i>Wine</i> . . . . .	61
6.1	Resultados para o classificador <i>1NN</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	69
6.2	Resultados para o classificador <i>10NN</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	70
6.3	Resultados para o classificador Discriminante Linear sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	71
6.4	Resultados para o classificador Árvore de Decisão sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	72
6.5	Resultados para o classificador Máquina de Vetor de Suporte sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	73

6.6	Resultados para o classificador <i>Multi Layer Perceptron</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	74
6.7	Resultados para o classificador <i>Bayesiano</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes). . . . .	75
6.8	Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados <i>Diabetes</i> . . . . .	76
6.9	Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados <i>Ionosphere</i> . . . . .	76
6.10	Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados <i>Well</i> . . . . .	77
6.11	Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados <i>Glass</i> . . . . .	77
6.12	Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados <i>Yeast</i> . . . . .	77
6.13	Resultados para o classificador <i>1NN</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	78
6.14	Resultados para o classificador <i>10NN</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	78
6.15	Resultados para o classificador Discriminante Linear sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	78
6.16	Resultados para o classificador Árvore de Decisão sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	79
6.17	Resultados para o classificador Máquina de Vetor de Suporte sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	79
6.18	Resultados para o classificador <i>Bayesiano</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	79
6.19	Resultados para o classificador <i>Multi Layer Perceptron</i> sem supressão e com supressão utilizando <i>Bagging</i> (30 pacotes) e SMOTE. . . . .	80
6.20	Tempo gasto para treinamento dos classificadores. . . . .	81
A.1	Distribuição das Bases de Dados. . . . .	84
A.2	Porcentagens de cada classe referentes à base de dados Câncer. . . . .	85
A.3	Porcentagens de cada classe referentes à base de dados Diabetes. . . . .	85
A.4	Porcentagens de cada classe referentes à base de dados <i>Monks-2</i> . . . . .	86
A.5	Porcentagens de cada classe referentes à base de dados <i>Monks-3</i> . . . . .	87
A.6	Porcentagens de cada classe referentes à base de dados <i>Ionosphere</i> . . . . .	87
A.7	Porcentagens de cada classe referentes à base de dados Íris. . . . .	88
A.8	Porcentagens de cada classe referentes à base de dados <i>Well</i> . . . . .	89
A.9	Porcentagens de cada classe referentes à base de dados Contraceptivos. . . . .	89

A.10	Porcentagens de cada classe referentes à base de dados <i>Wine</i> . . . . .	91
A.11	Porcentagens de cada classe referentes à base de dados <i>Glass</i> . . . . .	91
A.12	Porcentagens de cada classe referentes à base de dados <i>Yeast</i> . . . . .	92
A.13	Porcentagens de cada classe referentes à base de dados <i>Abalone</i> . . . . .	93

# Lista de Algoritmos

1	<i>CLONALG</i> . . . . .	38
2	<i>opt-aiNet</i> . . . . .	39
3	Algoritmo <i>Bootstrapping</i> . . . . .	47
4	Algoritmo <i>SMOTE(T, N, k)</i> . . . . .	49
5	Algoritmo <i>SUPRAIC</i> . . . . .	55
6	Algoritmo <i>SeleSup</i> . . . . .	66

# Capítulo 1

## Introdução

O reconhecimento é uma característica comum entre os seres vivos, desde uma bactéria que engloba uma biomolécula como possível fonte de alimento, até um mamífero que identifica uma fêmea fértil da sua espécie. Sem a capacidade de reconhecer, um indivíduo está mais sujeito, por exemplo, a ingerir uma substância danosa ao seu organismo, ou a não reproduzir. Igualmente, aqueles aptos a perceberem o que lhes é próprio do organismo e o que é estranho, tornam-se mais habilitados a tentar se defender contra invasões microbiológicas.

Quando um micróbio invade o organismo e é reconhecido pelo sistema imunológico, são disparados mecanismos de defesa que estimulam a produção de células brancas específicas para neutralizar tal patógeno. Há, portanto, um aumento exponencial dessas células que pode continuar durante dias. Quando a infecção é controlada, mecanismos de auto-regulação do sistema imunológico cuidam para que o excesso de células específicas para esse patógeno sejam eliminadas. Estes mecanismos são responsáveis, também, pela eliminação de células auto-reativas ou de baixa aptidão para agir como frente de defesa.

Partindo do estudo dos mecanismos envolvidos no processo de supressão e auto-regulação dos sistemas imunológicos, são propostos, neste trabalho, dois algoritmos imunológicos. Um com o propósito de classificação e outro para seleção de dados de treinamento.

O classificador proposto trabalha com a seleção de um subconjunto de treino significativo de forma que a construção do modelo final preditor é baseada apenas neste subconjunto. Como um aprimoramento desta técnica, concebeu-se um mecanismo para seleção de dados de treinamento.

Assim como certos atributos têm mais importância que outros no processo de classificação, há também exemplos com essas mesmas características. Há várias razões para que haja interesse na seleção de exemplos de treinamento que serão apresentados para um algoritmo classificador. Por exemplo, em casos onde há um número de registros suficiente, faz sentido aprender a partir daqueles mais relevantes, para aumento de desempenho computacional. Outro motivo é quando há um custo associado à definição dos

rótulos dos exemplos e há muitas instâncias disponíveis.

Neste trabalho, a seleção dos dados é feita por um mecanismo inspirado no comportamento dos linfonodos e células T com relação a anticorpos pouco adaptados, que são responsáveis por gerar estímulos de sobrevivência para os plasmócitos. O objetivo foi construir um mecanismo para ser hibridizado a métodos de classificação de modo que se aumente o desempenho na construção do modelo preditor, sem que isto acarrete em perdas na sua acurácia.

Este texto foi estruturado de forma que o segundo capítulo apresenta conceitos básicos sobre imunologia necessários para o entendimento das idéias contidas neste trabalho. O terceiro capítulo descreve as premissas principais envolvendo sistemas imunológicos artificiais bem como os algoritmos mais utilizados atualmente. Em seguida, no capítulo quatro fala-se sobre classificação e questões relacionadas a desbalanceamento de classes e seleção de exemplos de treinamento. No capítulo cinco é proposto um classificador inspirado nos mecanismos de auto-regulação dos sistemas imunológicos. Em seguida, no sexto capítulo é apresentado um mecanismo de seleção de dados de treinamento imuno-inspirado. O trabalho encerra com as considerações finais com relação a cada algoritmo, resultados, propostas de aplicações e continuação da pesquisa.

## Capítulo 2

# Conceitos Básicos sobre Imunologia

*Este capítulo apresenta os principais conceitos relativos aos sistemas imunitários biológicos, necessários para o entendimento dos capítulos seguintes.*

### 2.1 Introdução

A vida se baseia em um frágil equilíbrio de milhões de reações químicas simultâneas e eventos inerentes a um ser vivo, como as interações deste com o ambiente e outros seres.

Um acidente corriqueiro, uma pequena farpa que penetre a pele de um indivíduo cria uma instabilidade na homeostase do corpo. Neste caso, só existem duas possibilidades, ou o ser vivo reconhece que existiu uma agressão, o tipo de agressor e a melhor forma de combatê-lo, e de posse destas informações passa a atacá-lo e destruí-lo, sem destruir a si próprio; ou sucumbe. Ao aparato de reconhecimento, ação e controle do organismo frente às agressões bioquímicas dá-se o nome de Sistema Imunitário ou Imunológico <sup>1</sup>.

O Sistema Imunitário (SI) é hoje conhecido como um dos sistemas biológicos mais complexos. Apesar da especificação da sua função ser a princípio bastante simples, detectar e eliminar qualquer organismo estranho, a execução desta tarefa não é trivial. Ainda hoje não se sabe exatamente como o SI funciona, nem mesmo qual a dimensão do seu papel dentro das interações do organismo [28]. Entretanto, à luz do que atualmente é conhecido sobre a sua dinâmica, muito se tem feito na tentativa de construir sistemas artificiais inteligentes contendo algumas de suas propriedades [37]. Isto por duas razões: a primeira é tentar trazer para o computador partes do SI através de modelos mais rudimentares e simulações do seu funcionamento, no intuito de entender melhor as tarefas que ele executa. A segunda, utilizar o SI como inspiração para a construção de sistemas computacionais inteligentes.

Há muito interesse neste segundo tópico, pelo fato de estarem presentes naturalmente no SI propriedades buscadas pelos projetistas na construção de sistemas. Estas são robus-

---

<sup>1</sup>Também pode ser chamado de Sistema Imune.

tez, tolerância a erros, reconhecimento de padrões, distribuição, memória, adaptabilidade, auto-organização e autonomia.

Este capítulo tem como objetivo discutir os conceitos fundamentais, as principais propriedades do SI e os mecanismos mais importantes das respostas imunes, a fim de fornecer subsídios para que o leitor entenda melhor os modelos artificiais propostos neste trabalho.

## **2.2 Aspectos Gerais das Respostas Imunitárias**

Ao se tentar enxertar um pequeno fragmento de tecido vivo extraído de um animal em outro da mesma espécie, observa-se que em poucos dias este tecido será destruído e eliminado pelo organismo do indivíduo receptor. Este processo, conhecido como resposta imunitária, é caracterizado pela capacidade que o organismo possui de reconhecer e destruir material estranho, sem implicar uma conseqüência fisiológica ou patológica de tal reação. Ao estudo das respostas imunitárias, suas conseqüências e mecanismos dá-se o nome de Imunologia.

O exemplo de rejeição de tecido visto acima representa apenas uma das formas nas quais a resposta imunitária está presente. Entretanto, ele também é um indicador importante da existência de um sistema eficiente de detecção e eliminação de células e moléculas que se apresentam estruturalmente diferentes daquelas tidas como normais e pertencentes ao organismo. Isto quer dizer que, através da resposta imunitária decorrente do enxerto do caso anterior, é possível deduzir que existe uma espécie de "sistema de vigilância constante", responsável pela identificação e remoção de células anormais, que podem ser pertencentes a outros seres vivos, como bactérias, ou mesmo células do próprio organismo contendo má formação ou defeitos, como por exemplo, aquelas presentes em tumores. Quanto mais eficaz for este sistema de localização e posterior remoção de irregularidades dentro do organismo, maiores são as suas chances de sobrevivência às invasões de parasitas e microorganismos nocivos.

A identificação das funções defensivas do sistema imune precedeu a concepção de que também uma de suas incumbências seria a detecção de irregularidades nas próprias células do organismo. A explicação para este fato está no registro histórico da evolução da humanidade e o surgimento das epidemias.

A partir daí, novos estudiosos dos fenômenos imunes e vacinas foram surgindo. O volume de pesquisas envolvendo o SI cresceu, e a Imunologia Moderna deste século se ocupa não só da descoberta de novas vacinas, mas também do estudo dos mecanismos do Sistema Imunológico, seus componentes bioquímicos, genéticos e a compreensão das interações biológicas dos organismos, na tentativa de prevenir e determinar tratamentos eficazes para doenças causadas por agentes externos.

## 2.3 Imunidade Natural e Adquirida

Como foi visto nos exemplos anteriores, o sistema imunológico dispõe de meios para se adaptar e tentar se proteger de patógenos que se infiltram no organismo ao longo da existência do indivíduo. Entretanto, existem também mecanismos de defesa não condicionados a estímulos externos e que já fazem parte do organismo do indivíduo desde o seu nascimento. Esta é a chamada **imunidade natural** (nativa ou inata) e sua estrutura básica é inerente a toda uma espécie, e não a um exemplar apenas; é de caráter geral e proporciona um tipo de defesa amplo, porém, não discriminado.

Fazem parte desta classe as barreiras físico-químicas da pele, membranas, saliva, lágrima, mucosas, enzimas digestivas, células fagocitárias, eosinófilos, um subconjunto de linfócitos denominados *Natural Killer* - NK, ou células matadoras naturais, e várias moléculas originadas no sangue <sup>2</sup> [1].

O outro tipo de imunidade é aquele que confere ao sistema imunológico a robustez de se proteger especificamente contra substâncias estranhas, a chamada **imunidade específica** ou adquirida. Este mecanismo de defesa é induzido ou estimulado pela exposição a fatores estranhos ao organismo; é particular para cada tipo distinto de macromolécula e se ajusta em magnitude e capacidade defensiva de acordo com a forma e intensidade da infecção, para cada indivíduo da espécie.

Foi convenicionado que a imunologia se ocupa do estudo da imunidade específica, contudo, fica difícil separar as duas formas de imunidade, visto que a parte adquirida da imunidade nada mais é do que uma evolução do Sistema Imunitário Inato, que torna os mecanismos de defesa do organismo mais potentes frente à igual evolução de patógenos, e favorecer a preservação da espécie. Ao longo do capítulo, este conceito se tornará mais claro, pois haverá vários exemplos onde as tarefas efetoras são realizadas pelos dois sistemas em conjunto, ou separadamente, porém, de forma redundante, o que comprova que não há entre eles uma linha divisora.

Dentro da imunidade, tanto inata como específica, há duas classes distintas que se diferenciam pela forma como a resposta imune se dá. São elas a **resposta imunitária humoral** e a **mediada por células**. As seções seguintes explicam de forma sucinta o funcionamento de cada uma delas, dando ênfase no seu papel dentro da imunidade adquirida.

## 2.4 Resposta Imunitária Humoral

No século XI, o mundo inteiro se via tomado por uma grande infestação de varíola. O índice de pessoas mortas decorrente desta doença naquela época foi muito grande, principalmente entre crianças. Na China, curandeiros andarilhos começaram a observar que certos indivíduos afortunados conseguiam se recuperar de uma primeira contaminação

---

<sup>2</sup>Complemento do sangue ou sistema complementar.

com a varíola e se tornavam resistentes a um posterior ataque da doença. Assim, estes curandeiros adquiriram o hábito de deliberadamente infectar crianças com extratos de pústulas da varíola com o objetivo de que aquelas que conseguissem sobreviver estivessem protegidas futuramente, caso houvesse uma nova infecção. Esta era uma operação indubitavelmente arriscada, entretanto, naquela época era aceitável, dado o alto índice de mortalidade infantil. Com o passar do tempo e o acúmulo de experiências na técnica, foi descoberto que era possível obter reações menos drásticas se a inoculação fosse dada a partir de pústulas originárias de casos mais brandos da doença. Este procedimento ficou conhecido como *variolação* (ou inoculação) e, em pouco tempo, se disseminou em direção à Europa.

Cem anos antes das evidências da natureza dos vírus serem demonstradas, em 1798, o médico inglês Edward Jenner observou que os trabalhadores rurais que ordenhavam o gado geralmente contraíam uma forma branda de varíola, adquirida do próprio gado. Seguindo a sugestão de uma de suas pacientes, uma vendedora de leite, Jenner passou a utilizar como substrato para a variolação o extrato de pústulas bovinas - a *vaccínia* ou *cowpox*, e batizou este procedimento de vacinação. O sucesso foi tal que culminou na popularização do procedimento e posterior erradicação da doença no mundo. Tem-se aqui o marco inicial da Imunologia.

Quando Jenner introduziu a vacinação, nada sabia a respeito dos agentes infecciosos causadores das enfermidades. No século XIX Robert Koch conseguiu provar que as diversas doenças infecciosas eram causadas por diferentes microorganismos patogênicos. Hoje em dia sabe-se que existem cinco grandes classes de agentes causadores das patologias [17]: as bactérias, os fungos, os parasitas, os vírus, e mais recentemente, os prions<sup>3</sup>.

Por volta de 1900, o pesquisador Louis Pasteur com seus estudos sobre a cólera aviária (*Pasteurella multocida*) foi capaz de promover um melhor entendimento dos métodos propostos por Jenner. Pasteur tentou sem sucesso infectar galinhas com uma cultura já envelhecida do microorganismo causador da cólera em aves. Em uma segunda tentativa, já com uma cultura mais recente da *Pasteurella multocida*, Pasteur se deparou com o estranho fato de que as galinhas infectadas na experiência anterior sobreviveram ao novo contágio, o que supostamente deveria ser letal. Intrigado com os resultados obtidos, levantou a hipótese de uma possível ligação entre o ocorrido com as galinhas e o episódio do uso varíola bovina por Jenner.

Esta suspeita impeliu Pasteur a estudos cujas conclusões deram fundamento ao processo de imunização proposto por Jenner e o levaram a desenvolver a vacina contra doenças como a cólera aviária e a raiva. Ele entendeu que na vacinação, a exposição de um

---

<sup>3</sup>PRIONS - Pequenas partículas infecciosas, proteínicas, que resistem à inativação por procedimentos que modificam os ácidos nucléicos e que contém uma proteína celular isoforme anormal. Podem ser transmissíveis e causam doenças degenerativas do Sistema Nervoso Central em humanos (Kuru, Doença de Creutzfeldt-Jakob, Síndrome de Gerstmann-Straussler) ou em animais (scrapie, encefalopatia espongiforme bovina, etc).

animal a uma cepa não virulenta de um microorganismo causador de doenças pode futuramente protegê-lo contra uma infecção virulenta do mesmo microorganismo ou de algum outro semelhante.

Apesar do seu sucesso no desenvolvimento de vacinas, Pasteur sabia muito pouco sobre a dinâmica envolvida no processo de imunização. Foi somente dez anos mais tarde (1890) que Emil von Behring e Sibasaburo Kitasato demonstraram que o surgimento dos mecanismos de defesa induzidos por intermédio da vacinação estavam associados a fatores de proteção encontrados no soro sanguíneo dos indivíduos vacinados. Um exemplo desta formação de fatores defensivos está no processo de fabricação de soros antiofídicos.

Os antídotos para veneno de cobra são comumente obtidos do sangue de cavalos. O veneno é extraído da serpente e diluído para que fique mais fraco. O cavalo recebe em intervalos de dias, várias aplicações deste preparado, em concentrações cada vez mais altas, porém nunca atingindo a concentração máxima, o que poderia ser danoso ao animal. Após um período de aproximadamente seis semanas, o organismo do cavalo se torna resistente à peçonha devido à produção de agentes protetores [77]. Para a obtenção do soro antiofídico, uma parcela do sangue do cavalo é retirada e faz-se o seu fracionamento com a finalidade de isolar os agentes de defesa, ou o antídoto propriamente dito. Assim, quando este contraveneno for injetado em uma pessoa ou animal picado por cobra de veneno semelhante ao inoculado no cavalo, espera-se que esta pessoa ou animal se torne temporariamente resistente e não sucumba aos efeitos danosos causados pela substância peçonhenta.

Os elementos do soro sanguíneo que conferem resistência ao organismo a agentes ofensivos como venenos e toxinas são os chamados **anticorpos**, e sua produção é incitada pela substância danosa injetada, ou **antígeno**. Os anticorpos atuam reagindo quimicamente com os antígenos de forma a neutralizar o seu princípio ativo e impedir que ele cause mais malefícios ao organismo. Os anticorpos são antígeno-específicos. Para um determinado antígeno há um anticorpo produzido especificamente para se ligar a ele. Da mesma forma, quando o antígeno penetra o organismo, somente o anticorpo específico terá sua produção estimulada.

Através da análise sanguínea em intervalos de tempo, é possível determinar a cinética da resposta imune do cavalo. Após a primeira exposição do animal ao antígeno, observa-se um período de latência (janela imunológica ou *lag*) de vários dias, onde não há nenhum tipo de alteração no soro. Em seguida, surge uma alta concentração de anticorpos <sup>4</sup> que pode, no decorrer de alguns dias, atingir o seu grau máximo e cair rapidamente, o que caracteriza a chamada resposta imunitária primária. Se uma nova dose do veneno for injetada após algum tempo decorrido desde a primeira reação, observa-se que o período *lag* é bem mais curto, e a quantidade de anticorpos aumenta rapidamente, caindo de forma lenta.

---

<sup>4</sup>Quantidade determinada pela capacidade de neutralização de uma certa medida de antígenos.

Esta diferença entre as respostas primária e secundária indica a existência de um aparato produtor de anticorpos capaz de gravar, em uma forma de memória, exposições anteriores a antígenos, provendo o SI da capacidade de se proteger mais rapidamente quando re-exposto a um agente danoso. Esta memória, em certos casos, pode ter sua dimensão reduzida com o passar do tempo. Entretanto, ela continua conferindo um tempo de latência mais curto em reações imunes secundárias, quando contrastado com as primárias. A figura 2.1 mostra de forma detalhada o mecanismo destas respostas imunes, levando em consideração a quantidade de anticorpos presentes no sangue e os dias decorrentes das inoculações do antígeno.

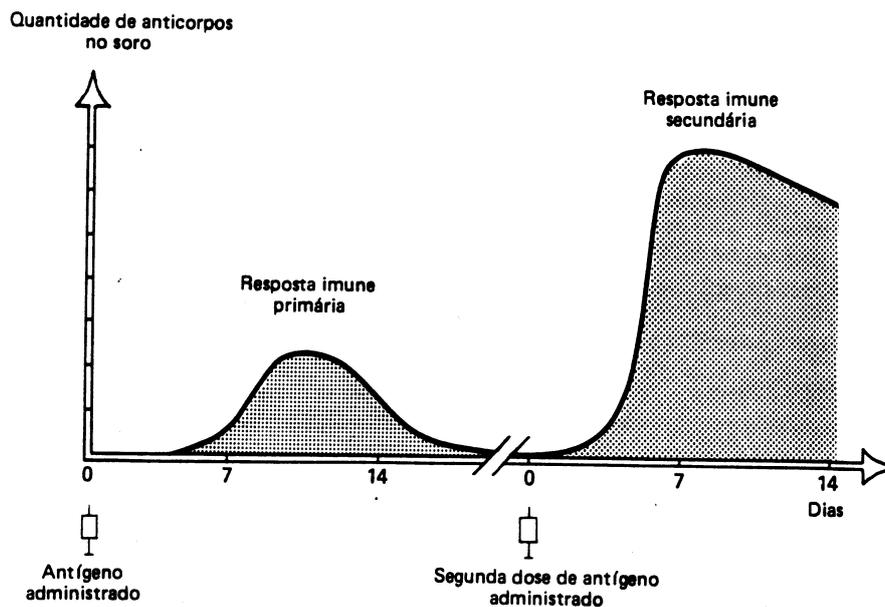


Figura 2.1: Cinética da resposta imunitária observando as concentrações de anticorpo no soro (adaptado de [77]).

## 2.5 Resposta Imunitária Mediada por Células

No caso do enxerto de tecido estranho da seção 2.2, estima-se que em cerca de dez dias esse tecido já terá sido completamente eliminado do organismo. Se, porventura, outro tecido com as mesmas características for novamente introduzido no corpo do indivíduo, a eliminação ocorrerá outra vez. Porém, o processo durará menos de dois dias. Vê-se que, neste novo exemplo, existem também mecanismos de respostas imunes primária e secundária. Contudo, o que há de diferente entre as respostas descritas previamente para o complexo humoral, é que agora quem participa principalmente destas respostas não são mais anticorpos, mas células. Destarte, a capacidade de um organismo reagir contra um tecido estranho, por exemplo, pode ser transmitida para outro através de células vivas chamadas linfócitos, que serão descritas com mais detalhes nas seções subseqüentes.

Um outro caso no qual a resposta celular pode ser encontrada é na defesa contra parasitas intracelulares, como os vírus e algumas bactérias. Os vírus são microorganismos compostos por ácido nucléico envolto por uma cápsula de proteína. Por não possuírem um aparato bioquímico necessário à reprodução, invadem células e assumem o comando, fazendo com que as células infectadas funcionem como seu aparelho reprodutor. A infecção viral altera o metabolismo da célula, podendo causar sua morte. Enquanto o vírus estiver no interior celular, está protegido pela membrana plasmática, o que impede às moléculas do complemento e os anticorpos atuarem contra ele. Para a defesa do organismo frente a este tipo de parasita, existem células imunocompetentes especializadas em vasculhar e destruir no organismo qualquer tipo celular do hospedeiro que apresente alguma irregularidade.

## 2.6 Características Gerais das Respostas Imunes

A partir do que foi mostrado a respeito das respostas imunitárias, é possível perceber algumas características importantes que devem estar presentes em um SI para que seu funcionamento seja considerado normal:

**Especificidade:** cada antígeno suscita no organismo um tipo específico de resposta;

**Diversidade:** estima-se que o sistema imune dos mamíferos seja capaz de discriminar aproximadamente  $10^9$  determinantes antigênicos distintos, ou seja, qualquer molécula estranha que penetre o organismo será identificada.

**Memória:** confere ao SI um desempenho melhor em suas respostas secundárias quando há uma re-exposição a um determinado antígeno.

**Tolerância:** capacidade do sistema em distinguir dentre as várias moléculas presentes dentro do indivíduo, o que faz parte do organismo e o que é externo.

**Auto-Regulação:** existem no SI mecanismos reguladores que proporcionam o equilíbrio com relação à ativação de agentes defensivos contra um determinado antígeno. Quando ocorre um ataque, há em um dado instante um aumento considerável dos fatores de defesa, seguido de redução à medida que o agressor é eliminado.

## 2.7 Mecanismo das Respostas Imunitárias

Em [77], há uma metáfora do funcionamento básico do sistema imune:

*“Em alguns sentidos, o sistema imune pode ser comparado a um estado totalitário em que os estrangeiros são expelidos, os cidadãos que se comportam são tolerados e aqueles que se “desviam” são eliminados. Ao mesmo tempo que esta analogia não deve*

*ser levada a extremos, torna-se rapidamente claro que tais regimes possuem vários aspectos característicos que incluem defesa de fronteira e uma força policial que mantém a população sob vigilância e elimina prontamente os dissidentes. As organizações dessa natureza também tendem a desenvolver um sistema de passaporte, de modo que os estrangeiros que não possuem certos aspectos de identificação são rapidamente detectados e remanejados”.*

Assim, quando um antígeno penetra o corpo, deve ser detectado, apreendido e reconhecido. Sendo identificado como estranho e danoso, acionam-se os sistemas imunitários para reagir por meio da produção de anticorpos ou células, eliminar o invasor, armazenando, também, informações sobre este evento a fim de que a resposta a uma nova invasão seja mais eficaz. Um esquema deste mecanismo pode ser visto na figura 2.2.

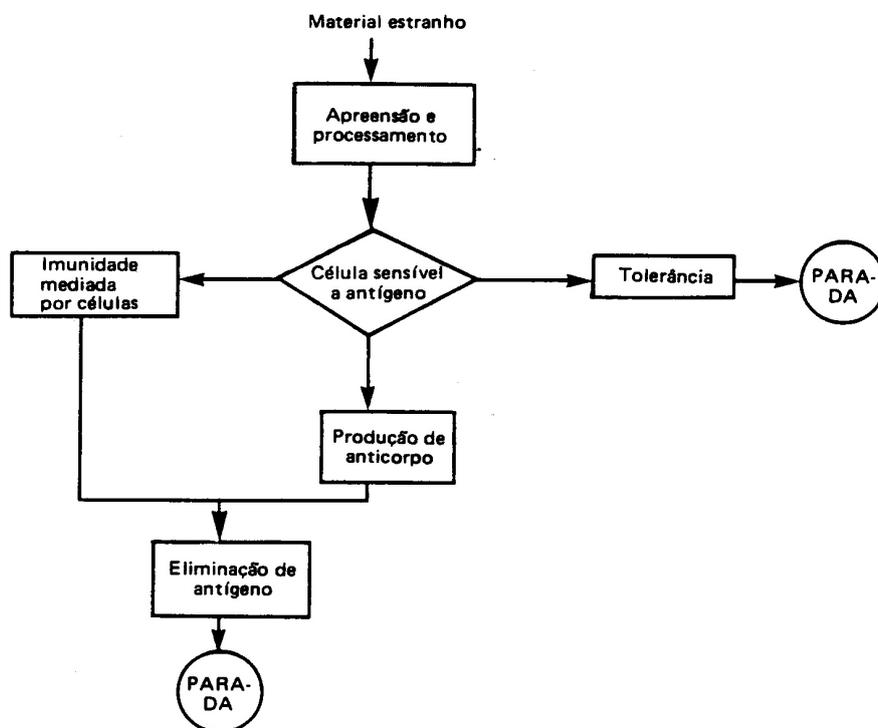


Figura 2.2: Aspectos essenciais das respostas imunitárias (adaptado de [77]).

Dentro do corpo, é possível distinguir os elementos que desempenham cada uma das funções citadas. A primeira linha de defesa celular do organismo é constituída por células responsáveis pela apreensão do patógeno, exposição e alerta de invasão. É constituída pelos macrófagos, granulócitos, células dendríticas e mastócitos. Os linfócitos são células sensíveis a antígenos, que efetuam o reconhecimento, participam das respostas imunitárias primária e secundária e trabalham como células efectoras nas respostas mediadas por células. As células responsáveis pela produção de anticorpos são derivadas dos linfócitos e conhecidas como plasmócitos. Cada um destes integrantes será visto com mais detalhes nas seções seguintes.

## 2.8 Antígeno e Antigenicidade

Um antígeno é qualquer molécula que pode ligar-se especificamente a um anticorpo. Seu nome deriva de sua propriedade de GERar ANTicorpos. Alguns antígenos, contudo, não o fazem por si mesmos, reservando-se a designação de imunógenos apenas para aqueles antígenos que podem induzir a produção de anticorpos.

Os fagócitos são células do sistema inato as quais trabalham capturando qualquer material estranho encontrado no organismo do hospedeiro. Entretanto, não se pode dizer que todo esse material apreendido desencadeia uma resposta imune adaptativa. Na verdade, para ser considerado antigênico, o corpo estranho deve possuir algumas características com relação à natureza físico-química e o material estranho que o compõe deve ser constituído de substâncias não toleradas pelo sistema imune do indivíduo infectado.

Sob o ponto de vista das limitações físico-químicas, para serem consideradas antigênicas, as moléculas devem ser grandes <sup>5</sup>, rígidas e quimicamente complexas [77]. Desta maneira, macromoléculas de estrutura complexa, como as proteínas, são consideradas melhores antígenos. Dentro de uma molécula de um antígeno, podem haver vários sítios de ligação denominados regiões antigênicas. Isto possibilita a ligação de mais de um anticorpo específico em uma molécula. Se existe em um patógeno uma região comum ou semelhante a uma outra molécula já detectada e eliminada pelo SI, os anticorpos já existentes no sangue serão capazes de se ligar a esta nova molécula. A este fenômeno se dá o nome de reação cruzada.

Uma outra característica relacionada às restrições físico-químicas é a degradabilidade. Moléculas que não podem ser degradadas e processadas pelos fagócitos desencadeiam pouca ou nenhuma resposta imune. Um exemplo prático descrito em [77] é o uso de válvulas cardíacas de porco fixadas com glutaraldeído na cirurgia cardíaca humana. O glutaraldeído fixa as proteínas da válvula, tornando-as metabolicamente inertes e não antigênicas.

A segunda característica para que uma substância seja antigênica é o não reconhecimento pelas células sensíveis a antígenos. Ela define quais moléculas podem ser consideradas próprias e podem permanecer no organismo e quais devem ser eliminadas. Para que não haja reação de linfócitos com auto-antígenos, há mecanismos que os eliminam antes de sua completa formação. Quando há falhas nos mecanismos de eliminação de células que atacariam o que é próprio do organismo, têm-se as chamadas doenças auto-imunes, como a artrite reumatóide, lúpus eritematoso sistêmico e diabetes *mellitus* tipo I.

---

<sup>5</sup>Há moléculas pequenas que atuam como antígenos, porém as maiores são mais eficazes para este fim.

## 2.9 Células e Tecidos do Sistema Imune

Todas as células do organismo, inclusive as células brancas, têm sua origem em um único grupo de células progenitoras, denominadas células-tronco hematopoiéticas da medula [38], como pode ser visto no esquema da figura 2.3. Neste texto, há interesse em estudar os derivados dos progenitores mielóide e linfóide comum que são as células que se diferenciam em leucócitos dos sistemas inato e adaptativo.

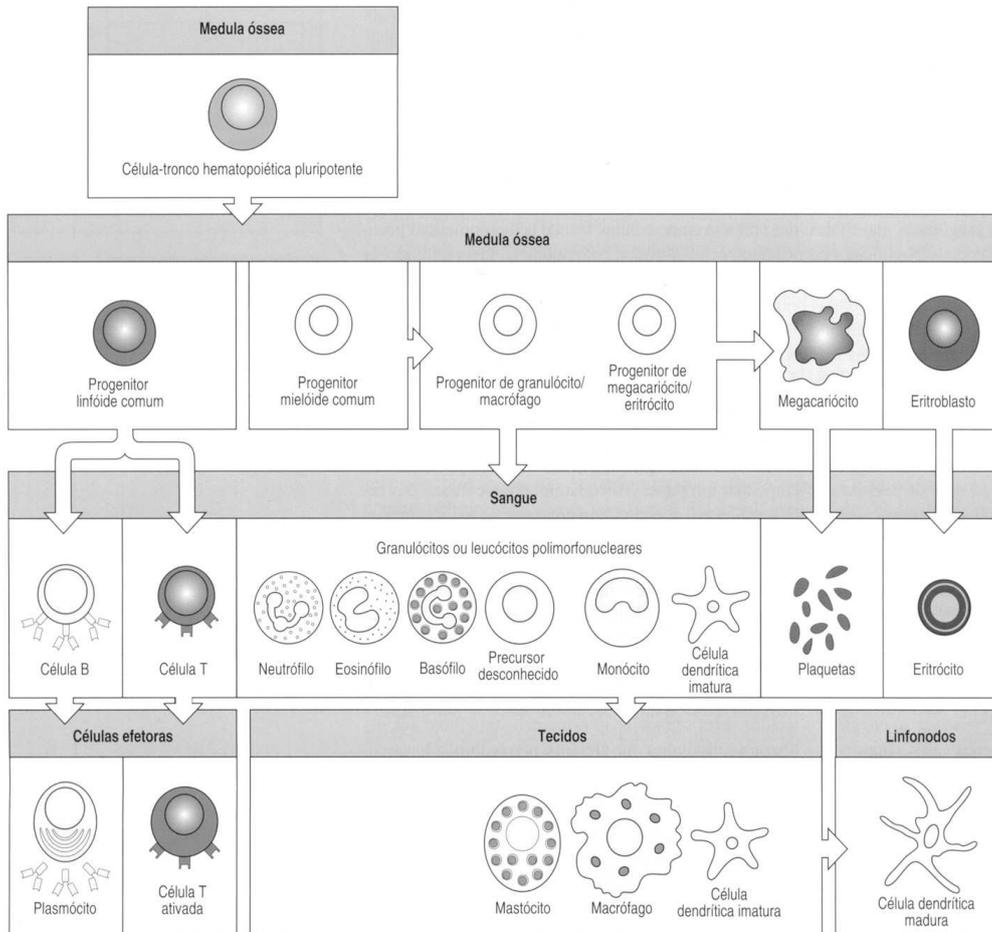


Figura 2.3: Hematopoiese (adaptado de [38]).

O progenitor mielóide é o precursor dos granulócitos, macrófagos, células dendríticas e mastócitos do sistema imunitário [38]. Os macrófagos constituem um dos três tipos de fagócitos presentes no SI. Este tipo de célula se encontra amplamente distribuída no sangue e tecidos. Em sua forma imatura, é conhecido como monócito, tornando-se macrófago após a sua migração da corrente sanguínea para os tecidos. Representa uma das células mais importantes da imunidade inata.

O papel das células dendríticas é capturar e apresentar o antígeno aos linfócitos, a fim de que ele seja reconhecido e combatido. Realizam tanto a fagocitose como a micropinocitose. Existem em sua superfície receptores capazes de reconhecer estruturas

comuns a diversos tipos de patógenos. Assim, quando esses receptores encontram uma molécula patogênica reconhecível, a célula dendrídica é estimulada a englobar o patógeno e degradá-lo. Ao ingerir o patógeno, a célula se ativa, amadurecendo em uma célula apresentadora de antígenos (APC - *Antigen Presenting Cell*) e migra até o linfonodo mais próximo. Já amadurecida, a célula é capaz de ativar os linfócitos antígeno-específicos do linfonodo e secretar citocinas que determinam quando e como o sistema imune deverá responder aos agentes infecciosos.

Os mastócitos influenciam na permeabilidade vascular, gerenciam as respostas alérgicas e acredita-se que estejam também envolvidos na proteção da superfície de mucosas [38].

Os granulócitos, ou leucócitos polimorfonucleares recebem essa denominação por conterem grânulos altamente coráveis em seu citoplasma. Há três tipos de granulócitos. Em todos os seus tipos, apresentam vida média curta. A sua produção é mais estimulada durante as respostas imunes, onde migram do sangue para os locais de infecção ou inflamação. Os neutrófilos são as células fagocitárias mais numerosas e importantes do complexo inato. Os eosinófilos estão presentes nas respostas contra infecções parasitárias. Os basófilos possuem funções ainda indeterminadas; entretanto, acredita-se que atuam auxiliando as respostas realizadas pelos mastócitos e eosinófilos. Um quadro com o resumo das células descritas e suas funções pode ser visto na figura 2.4.

O progenitor linfóide comum dará origem aos linfócitos, que serão descritos mais detalhadamente no próximo tópico e ao longo do texto.

### **2.9.1 O Tecido Linfóide**

O tecido linfóide é constituído por um conjunto de órgãos e tecidos composto pelo timo, os nódulos linfáticos (folículos linfóides), os linfonodos e o baço (figura 2.5). Em todas as suas formas, está presente neste tecido uma população densa de linfócitos [10], o que pode ser facilmente explicado pelo papel dessas células nas respostas imunes.

### **2.9.2 Os Órgãos Linfóides**

Os órgãos linfóides são tecidos carregados de linfócitos em um ambiente de células não-linfóides [38]. Nestes órgãos, o contato dos linfócitos com células não-linfóides proporcionam seu desenvolvimento e muitas vezes iniciam respostas imunes adaptativas. São divididos em órgãos centrais – ou primários – e periféricos.

Os órgãos centrais compreendem a medula óssea e o timo. Recebem essa denominação por constituírem ambientes de geração de linfócitos. Os órgãos periféricos são assim chamados por serem os locais onde as respostas adaptativas se iniciam e onde os linfócitos se alojam. São compostos pelos linfonodos e baço. O esquema da figura 2.5 mostra em detalhes os órgãos linfóides e sua localização no corpo humano.

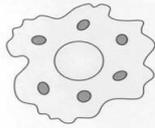
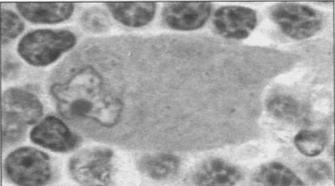
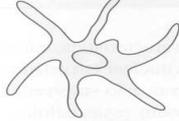
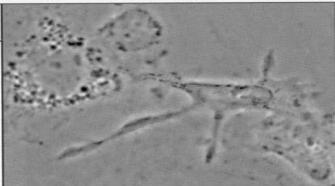
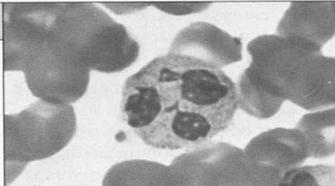
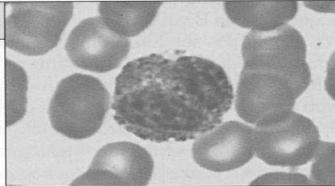
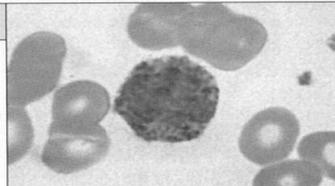
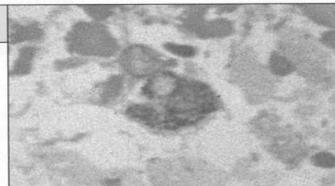
Células		Função ativada
Macrófago	 	<p>Fagocitose e ativação de mecanismos bactericidas</p> <p>Apresentação de antígeno</p>
Célula dendrítica	 	<p>Captura do antígeno nos locais periféricos</p> <p>Apresentação de antígenos nos linfonodos</p>
Neutrófilo	 	<p>Fagocitose e ativação de mecanismos bactericidas</p>
Eosinófilo	 	<p>Morte de parasitas recobertos por anticorpos</p>
Basófilo	 	<p>Desconhecida</p>
Mastócito	 	<p>Liberação de grânulos contendo histamina e outros agentes ativos</p>

Figura 2.4: Células do SI inato e suas funções (adaptado de [38]).

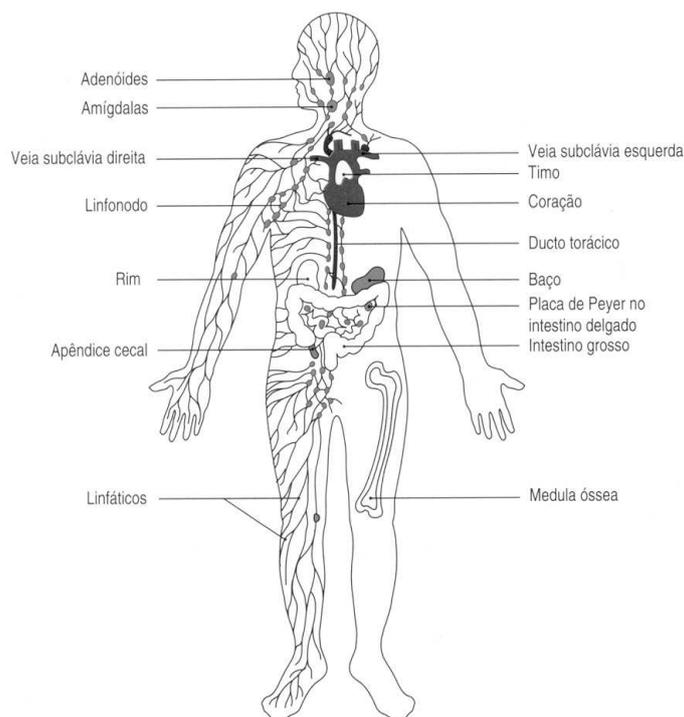


Figura 2.5: Distribuição dos tecidos linfóides (adaptado de [38]).

### 2.9.3 Órgãos Linfóides Primários

#### Medula Óssea

A medula óssea se localiza no interior dos ossos longos e é constituída por um tecido esponjoso mole. Ela é responsável pela hematopoiese, produzindo os leucócitos, hemácias e plaquetas do sangue. É o principal órgão renovador das células do SI.

#### O Timo

O timo é um órgão localizado no espaço mediastinal anterior [77] e cujo tamanho pode variar consideravelmente entre os mamíferos. É o local onde os linfócitos T são gerados ou amadurecem, quando têm origem na medula. O timo atinge o seu tamanho máximo na puberdade, entretanto, sua atividade é mais intensa na vida fetal e primeira infância, onde os clones das células T estão sendo formados. Após a puberdade, ele sofre atrofia e a produção dos linfócitos diminui. Existe nas aves um órgão equivalente ao timo que serve para maturação de células B. Este órgão é chamado *Bursa de Fabricius*.

## 2.9.4 Órgãos Linfóides Secundários

### Os Linfonodos

Os linfonodos são estruturas linfóides altamente organizadas, localizados em pontos de convergência do sistema linfático [38]. Como pode ser visto na figura 2.6, existem locais específicos para cada tipo de linfócito dentro do linfonodo. Os linfócitos B se concentram em folículos e os linfócitos T ocupam uma área mais vasta, constituindo as chamadas zonas de células T. Alguns folículos possuem pontos centrais denominados centros germinativos, que são caracterizados por intensa proliferação de linfócitos B, quando há encontro com antígenos.

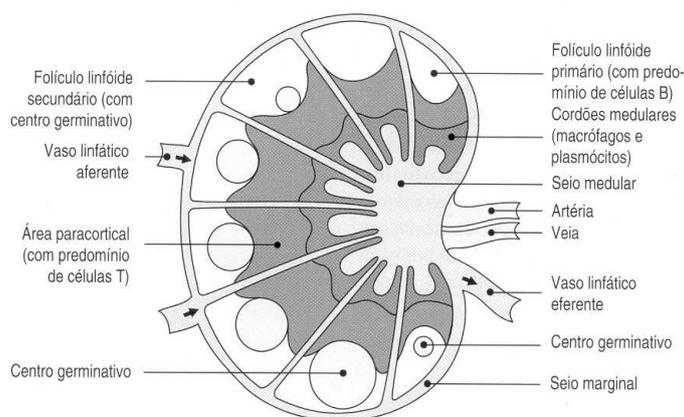


Figura 2.6: Esquema de um linfonodo (adaptado de [38]).

Pode-se dizer que os linfonodos atuam como “centros de inteligência” do SI, onde é procurado qual o linfócito mais apto a combater o antígeno apresentado pela APC e se dá o início à resposta ao agressor. Contudo, o propósito dos linfonodos vai além das tarefas de alojar linfócitos e facilitar o mecanismo de defesa adaptativo. Eles, assim como todos os demais constituintes dos tecidos linfóides periféricos, são responsáveis por enviar continuamente sinais de sobrevivência para os linfócitos que ainda não encontraram seu antígeno específico [38].

Existem outros agregados de células linfóides espalhados no organismo. Incluem tecidos linfóides relacionados ao intestino ou GALT – *gut-associated lymphoid tissues* – constituídos pelas tonsilas, adenóides, apêndice cecal e as placas de Peyer responsáveis por coletar antígenos das superfícies epiteliais do trato gastrointestinal [38]; BALT – *bronchial-associated lymphoid tissue* – que protegem o sistema respiratório e MALT – *mucosal associated lymphoid tissues* – relacionado à proteção das mucosas.

### O Baço

O baço é o maior órgão linfóide secundário. Sua função é parecida com a dos linfonodos, com a diferença que o baço não possui ligação direta com os vasos linfáticos, mas com

a corrente sanguínea. Ele é, portanto, o local onde os linfócitos combatem os agentes infecciosos do sangue. A figura 2.7 mostra um esquema do baço contendo a organização dos seus tecidos linfóides. Ele é constituído pela polpa branca que responde aos antígenos levados pelo sangue e pela polpa vermelha, que remove células senescentes.

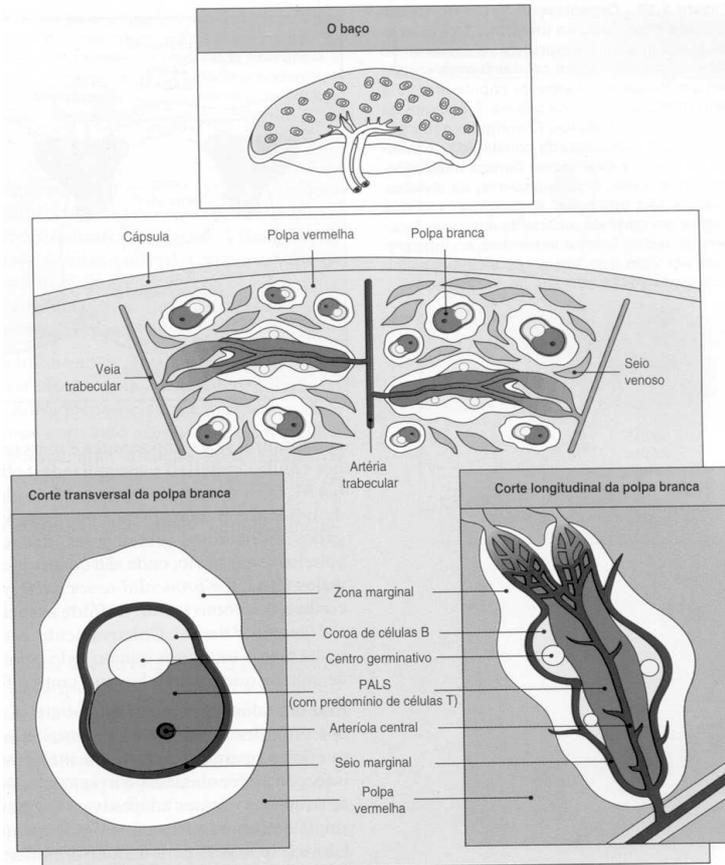


Figura 2.7: O baço (adaptado de [38]).

### 2.9.5 Os Linfócitos

Os linfócitos são células imunocompetentes recirculantes responsáveis pela imunidade específica. Estas células realizam um intercurso contínuo entre corrente sanguínea e linfa, e em seguida no sentido inverso, num processo chamado recirculação [10]. A linfa origina-se do líquido intersticial excedente que os capilares sanguíneos são incapazes de reabsorver, como pode ser visto na figura 2.8. Este líquido, antes de retornar para a corrente sanguínea, é absorvido pelos capilares linfáticos e drenado ao longo dos vasos linfáticos até os linfonodos. É nos linfonodos que haverá uma triagem determinando quais componentes da linfa poderão retornar ao sangue.

Assim, um linfonodo situado ao longo de um vaso linfático é capaz de filtrar diversos tipos de antígenos e células estranhas que possam estar presentes na linfa, evitando que estes patógenos penetrem na circulação sanguínea. Do mesmo modo, o baço é res-

ponsável por filtrar o sangue. Cabe aqui ressaltar que a intensidade desta barreira imposta pelo posicionamento dos linfonodos e do baço depende da sua população residente de macrófagos.

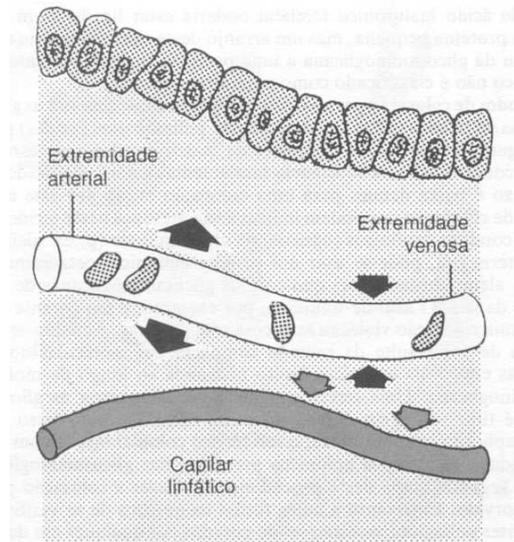


Figura 2.8: Formação do líquido tecidual. (adaptado de [10]).

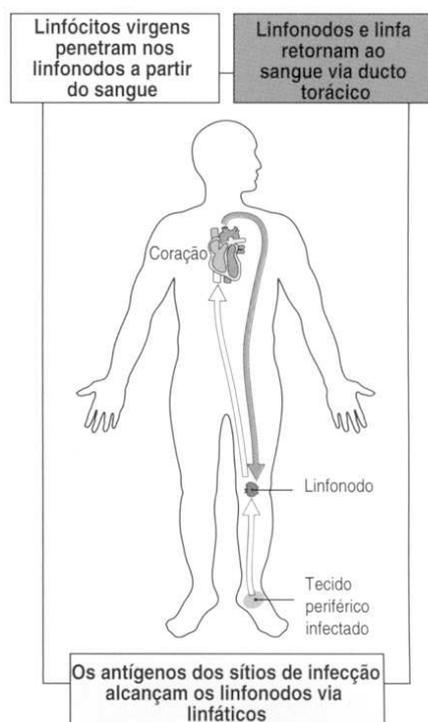


Figura 2.9: O caminho percorrido pelos linfócitos e seu encontro com os antígenos. (adaptado de [38]).

Durante o movimento de recirculação do sangue para a linfa e novamente para a corrente sanguínea, ilustrado na figura 2.9, há grandes chances dos linfócitos encontrarem

antígenos que possam ter penetrado o organismo. Um encontro de tal natureza pode, muitas vezes, desencadear uma resposta imune, pois esse tipo de célula, em sua maioria, dispõe da capacidade funcional para reconhecer e responder a patógenos. Quando um pequeno linfócito reconhece um antígeno, ele aumenta de tamanho, replica o seu DNA e passa por uma série de divisões dando origem aos chamados clones, que são conjuntos de células identicamente programadas para o combate a um determinado agente agressor [10].

É possível destacar duas classes principais de linfócitos. Aqueles cuja diferenciação se dá no Timo são chamados linfócitos T e os outros cujo desenvolvimento se dá na medula óssea (*Bone Marrow*) – são os linfócitos B.

## Os Linfócitos T

Os linfócitos T representam a maioria dos linfócitos do sangue e são os principais responsáveis pela imunidade adaptativa. Assim como os linfócitos B, eles tem sua origem na medula óssea, entretanto, sua fase de maturação se dá no Timo. É nesta fase de maturação que o linfócito terá a sua função e especificidade antigênica definidos. Apesar do seu aspecto microscópico de maneira geral ser uniforme, os linfócitos T se subdividem em diversas classes funcionais distintas [10], das quais se destacam as células T regulatórias e as células T citotóxicas ou citolíticas.

Há dois subtipos de células T regulatórias: as células T auxiliares (*T helper*) e as células T supressoras. As células T auxiliares são responsáveis pela secreção de linfocinas<sup>6</sup> que desencadeiam a ativação e transformação das células B "virgens" em plasmócitos. Atuam também na ativação das células T citotóxicas e de leucócitos inflamatórios [1]. Os linfócitos T auxiliares são as principais células infectadas pelo vírus da imunodeficiência adquirida (AIDS) [58]. Ao segundo subtipo de células regulatórias cabe a tarefa de inibir respostas imunes, ou mesmo suprimir a função de uma célula B diretamente. Ela também é capaz de tolher a ativação de uma célula T auxiliar, podendo ser, em certos casos, antígeno-específica.

As células T citotóxicas são as chamadas matadoras naturais e protagonizam a resposta imune mediada por células. Elas atuam lisando células estranhas ou aquelas infectadas por vírus e outros patógenos intracelulares.

A identificação destas subclasses de linfócitos T se deu com a descoberta de uma relação existente entre o papel realizado por uma célula T e certas proteínas expressadas em sua membrana. Por exemplo, constatou-se que a maioria das células T auxiliares expressa uma proteína de superfície designada como CD4<sup>7</sup>. Já nas células citolíticas, a proteína que se destaca é a CD8 [1]. Com isto, puderam ser identificados diversos sub-

---

<sup>6</sup>Linfocinas são fatores solúveis semelhantes a hormônios que afetam outras células (do grego *knesis*, movimento).

<sup>7</sup>CD significa *cluster differentiation*.

grupos específicos de linfócitos diferentes das células B e T. Um terceiro grupo de células foi destacado por não possuir marcadores de membrana com características específicas que os identifique como linfócitos B ou T. Ele é composto pelas chamadas células nulas, células *natural killer* e células *killer* dependentes de anticorpos. Há suspeitas que nelas se incluem certos estágios iniciais de diferenciação dos linfócitos B e T, entretanto, sua ontogênese e especificidade ainda não estão completamente elucidados. Sabe-se, contudo, que sua função envolve atividade citolítica dirigida contra alguns tipos de células neoplásicas, certas células enxertadas e alguns microorganismos.

### O Reconhecimento dos Antígenos pelas Células T

Os locais de reconhecimento de antígenos dos linfócitos T, ou receptores das células T, conforme a ilustração da figura 2.10, são formados por duas cadeias polipeptídicas  $\alpha$  e  $\beta$ , cada qual com uma região variável e uma constante ligadas por uma ponte dissulfídrica e pequenas cadeias laterais de oligossacarídeos firmemente presas na membrana celular [10, 38].

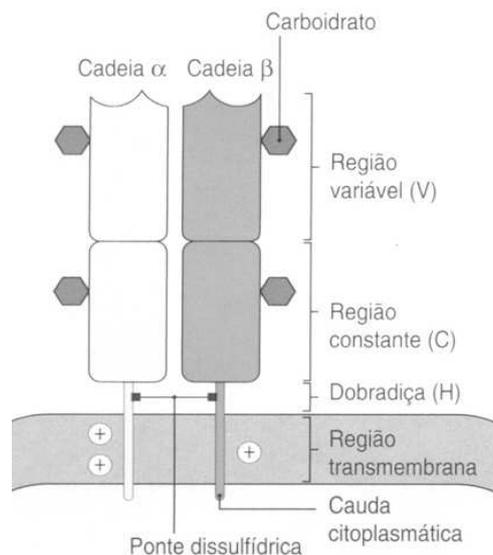


Figura 2.10: Esquema de um TCR (adaptado de [38]).

A figura 2.11 mostra a semelhança existente entre os receptores da célula T (TCR) e as regiões Fab das Imunoglobulinas, que serão vistas nas seções seguintes. Há um outro tipo de receptor, constituído por cadeias polipeptídicas diferentes  $\gamma$  e  $\delta$ , que parece possuir propriedades diferentes de reconhecimento do antígeno em relação aos receptores  $\alpha:\beta$ . Sua função, contudo, ainda não é totalmente conhecida. A especificidade de cada célula tem origem nas diversas combinações genéticas que codificam as regiões variáveis do receptor.

O processo de reconhecimento de um antígeno é semelhante ao que acontece com os linfócitos B, entretanto, as células T não são capazes de identificar um antígeno sozinhas.

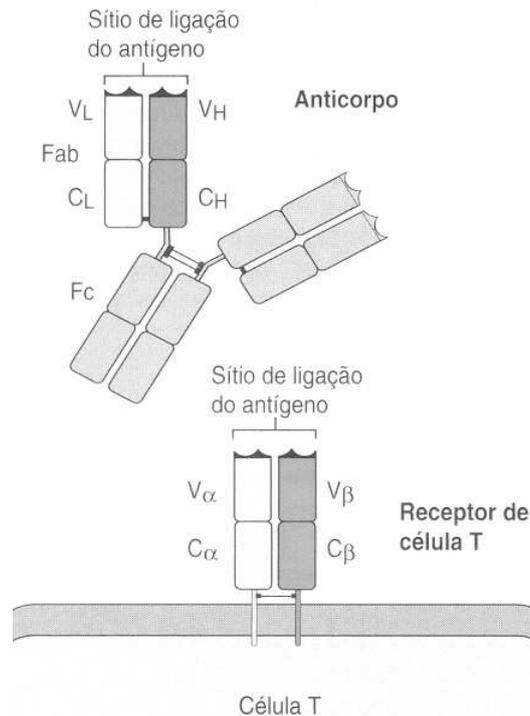


Figura 2.11: Semelhança entre os TCR e as regiões Fab das Imunoglobulinas (adaptado de [38]).

A maioria das células T só responderá a um antígeno estranho se este lhes for apresentado por uma célula acessória, que pode ser um macrófago ou uma célula apresentadora de antígenos, juntamente com as glicoproteínas integrais do Complexo Maior de Histo-compatibilidade (MHC) [10]. O MHC é o conjunto de genes responsável por codificar estas proteínas. Este mecanismo de apresentação de um antígeno, associado a partes integrantes do próprio organismo, confere ao linfócito T a capacidade de identificar células do corpo que contenham alguma irregularidade, como a expressão de antígenos estranhos em sua membrana (por exemplo, células infectadas por parasitas intracelulares) ou até mesmo células que apresentem glicoproteínas de um MHC estranho.

Há duas classes de moléculas de MHC. Elas são denominadas MHC classe-I e MHC classe-II. As moléculas da primeira classe são encontradas em todas as células, enquanto que as da segunda estão presentes nas células apresentadoras de antígenos [16]. As células T citotóxicas respondem a antígenos ligados a moléculas do MHC classe-I e as T auxiliares interagem com antígenos ligados a moléculas do MHC classe-II.

### Os Linfócitos B

Como já foi dito anteriormente, cabe aos linfócitos B a produção de anticorpos (ou imunoglobulinas), sendo que cada clone de linfócitos B é responsável por um antígeno em específico. Antes de sua ativação, cada linfócito B apresenta em sua superfície pequenas áreas de imunoglobulinas específicas, também chamadas de parátomos (em sua maioria,

sIgM e sIgD das classes IgM e IgD). Estas áreas são responsáveis pelo reconhecimento do antígeno através da identificação do seu determinante antigênico, ou epítopo, que são as áreas contra as quais a resposta imunitária tende a se dirigir e onde os anticorpos costumam se ligar [77]. Assim, quando um linfócito B encontra e reconhece um antígeno, ele se torna ativado.

Durante o seu processo de ativação, o linfócito aumenta <sup>8</sup>, reproduz e sua progênie se diferencia em células secretoras de anticorpos, os plasmócitos, e algumas células de memória de mesma especificidade antigênica determinada pelas imunoglobulinas da superfície celular do primeiro linfócito.

Enquanto a função dos plasmócitos é basicamente a liberação de anticorpos, as células de memória adquirem mais imunoglobulina de superfície (sIg) durante a sua diferenciação, mantendo o aspecto morfológico dos pequenos linfócitos e com um ciclo de vida muito maior que os plasmócitos - elas podem permanecer no organismo durante vários anos. Estes linfócitos B de memória são os responsáveis por uma resposta humoral secundária mais imediata e mais ampla.

A explicação para que a capacidade de reconhecimento dos linfócitos B seja tão abrangente, cobrindo praticamente todos os antígenos existentes, está na forma como é estruturada uma molécula de anticorpo.

## 2.10 O Anticorpo

O anticorpo é o receptor da célula B em sua forma solúvel. A principal diferença entre os receptores da membrana dos linfócitos B e os anticorpos está na cadeia pesada. Nos receptores das células B a porção C-terminal da cadeia pesada apresenta uma seqüência hidrofóbica, ao passo que nas imunoglobulinas essa seqüência é hidrofílica, fato que permite a sua secreção.

Uma molécula de imunoglobulina é constituída de quatro cadeias de polipeptídios com pequeno número de oligossacarídeos ligados [10]. Destas quatro cadeias, duas são cadeias pesadas idênticas entre si, e as outras duas são cadeias leves também iguais entre si. As cadeias pesadas possuem esse nome por apresentarem maior número de aminoácidos que as cadeias leves. A forma da molécula pode ser vista na figura 2.12. Seu formato é de Y. Sua base é composta por partes das duas cadeias pesadas enquanto os braços são constituídos de parte restante das cadeias pesadas juntamente com uma cadeia leve. O local de reconhecimento dos antígenos, específico para cada tipo diferente de patógeno, fica localizado na ponta de cada um dos braços do Y e é denominado região variável, ou região V. Os demais componentes da molécula constituem a região constante C. A especificidade do sítio de reconhecimento dos antígenos é determinada pela sua seqüência de

---

<sup>8</sup>Nesta fase ele recebe o nome de linfoblasto.

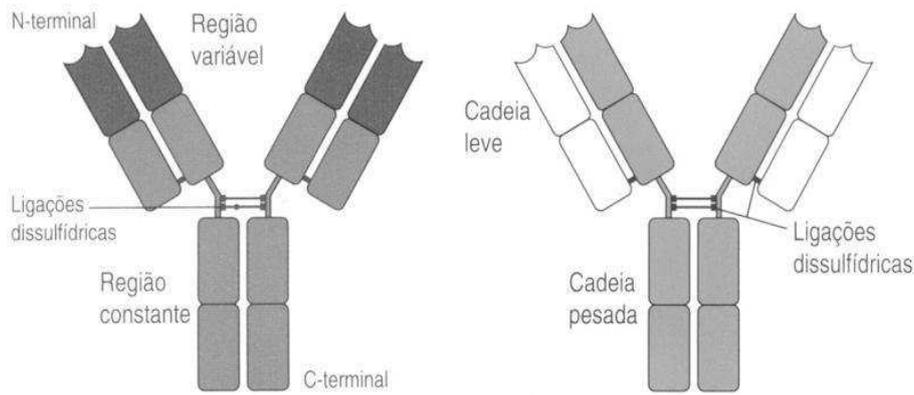


Figura 2.12: O anticorpo considerando sua divisão em regiões (esquerda) e cadeias (direita) (adaptado de [38]). Cada braço do Y é formado pela associação de uma cadeia leve com a metade amino-terminal de uma cadeia pesada, enquanto a perna do Y é formada pelo pareamento das metades carboxi-terminais das duas cadeias pesadas [38].

aminoácidos, e a sua diversidade é conferida pelos genes codificadores dessas seqüências e pelas mutações somáticas, o que será discutido com mais detalhes na seção 2.11.

A função de reconhecimento e neutralização do antígeno fica a cargo da região V. A região C participa na ativação de mecanismos efetores e, diferente da região V, a região C possui um número bastante reduzido de possíveis configurações estruturais. Dessa maneira, é possível organizar as imunoglobulinas (Igs) em cinco classes principais, de acordo com a forma da região C ou isotipo: IgA, IgD, IgE, IgG e IgM. As peculiaridades de cada classe de molécula estão expostas na figura 2.13. A IgG é a imunoglobulina de maior concentração no soro sanguíneo.

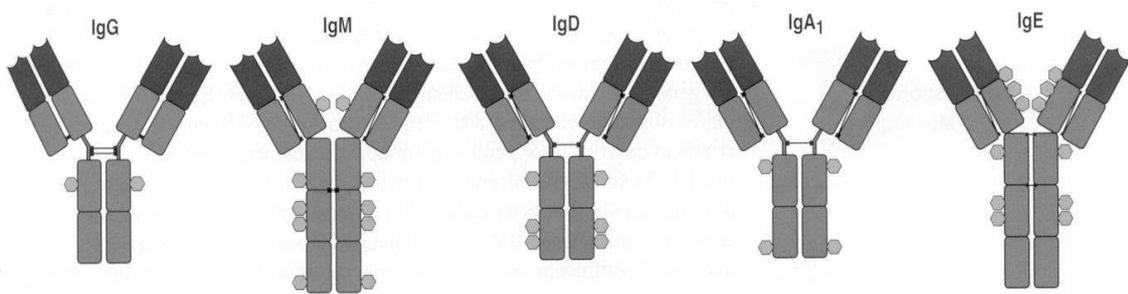


Figura 2.13: Organização estrutural das principais classes das Igs (adaptado de [38]).

### 2.10.1 A Interação da Molécula do Anticorpo com o Antígeno Específico

O tipo de força de ligação entre antígeno e anticorpo envolve interações não-covalentes. Em sistemas biológicos, as ligações não-covalentes permitem que haja uma formação rápida de complexos, com flexibilidade e de forma reversível. Isto possibilita a reutilização

dos anticorpos e sua adaptação à superfície da molécula antigênica. As ligações não-covalentes se formam com a proximidade das moléculas, assim, quanto mais ajustado um anticorpo estiver de um antígeno, maior a força de ligação entre eles. As interações não covalentes conhecidas entre antígeno e anticorpo são ligação iônica ou eletrostática, pontes de hidrogênio, interações hidrofóbicas e forças de Van der Waals. A ocorrência de cada uma dessas forças dependerá da estrutura das moléculas envolvidas na ligação.

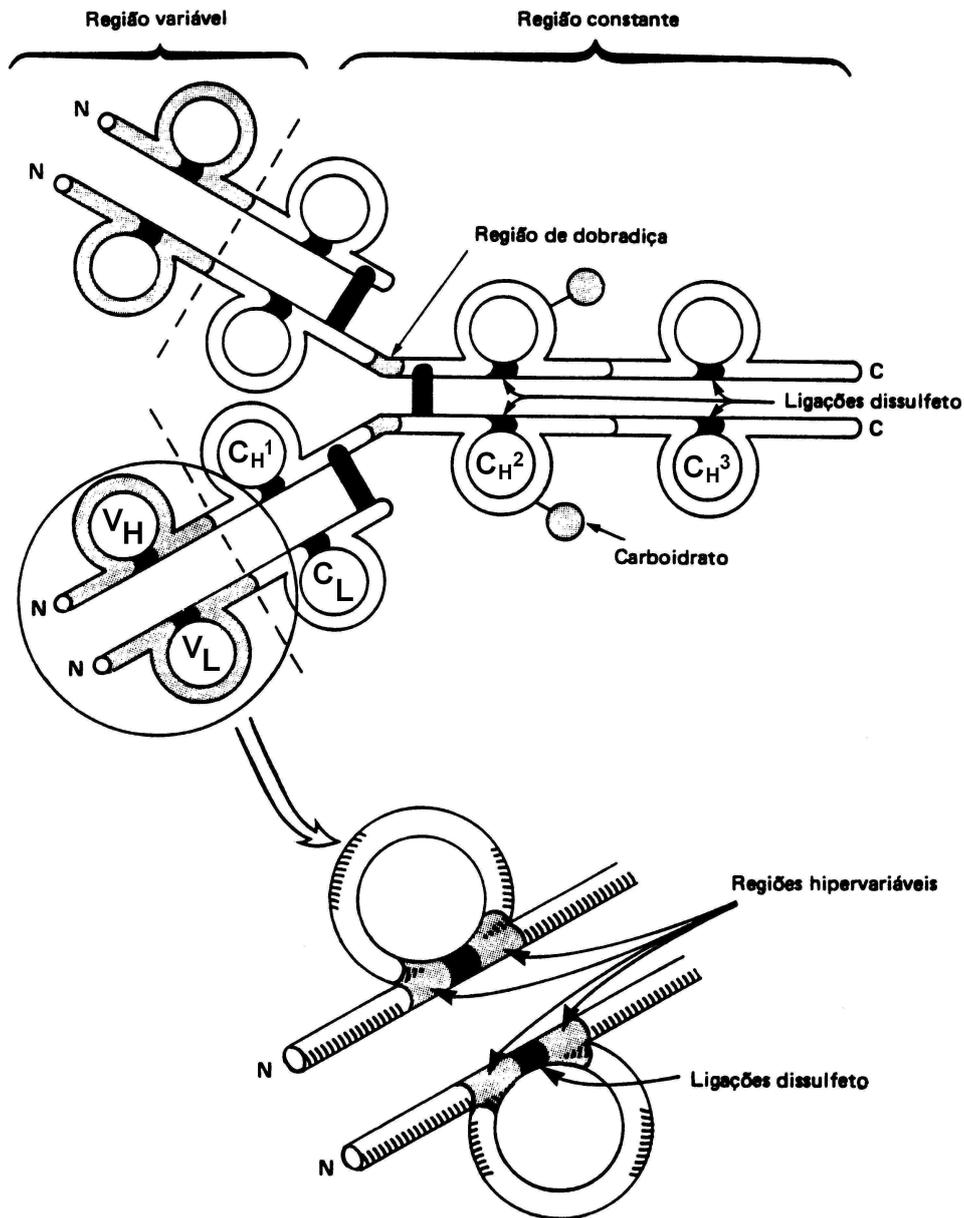


Figura 2.14: O anticorpo (adaptado de [77]).

### 2.10.2 Região Variável

É constituída de mais ou menos a metade da cadeia leve e um quarto da cadeia pesada. Existem ao longo desta região algumas áreas que possuem uma variabilidade muito maior

que as demais. Estas áreas são chamadas de regiões hipervariáveis. Este conjugado de regiões hipervariáveis e regiões de menor variação constituem o sítio de ligação do anticorpo.

### 2.10.3 Região de Dobradiça

Como pode ser visto no esquema da figura 2.14, existe uma região no anticorpo que permite que os braços do Y <sup>9</sup> de uma IgG se dobrem e girem livremente em torno do centro da molécula. Esta flexibilidade favorece a união dos braços do Y contra sítios de ligação de uma molécula de antígeno separados por distâncias diferentes.

### 2.10.4 Regiões Constantes

São formadas pela metade C-terminal de cada cadeia leve e dos três quartos finais de cada cadeia pesada [77]. A porção constante da cadeia leve é denominada  $C_L$ . A cadeia pesada é dividida em três subunidades similares, ou regiões homólogas  $C_H^1$ ,  $C_H^2$  e  $C_H^3$  tomando como base IgG <sup>10</sup>.

#### Região $C_L$

As cadeias leves são divididas em dois grupos  $\kappa$  e  $\lambda$ . Esta divisão se dá com base na antigenicidade e na seqüência de aminoácidos. As duas cadeias leves de uma imunoglobulina devem pertencer ao mesmo grupo.

#### Região $C_H$

A ativação dos mecanismos efetores fica a cargo da região constante. Certos sítios existentes na região constante das cadeias pesadas medeiam a ativação do complemento e a formação de imunocomplexos como preparação para ingestão por células fagocitárias [77].

## 2.11 A Geração da Diversidade dos Anticorpos

Antes da descoberta do mecanismo gerador da diversidade dos anticorpos, havia duas hipóteses principais para explicar a origem dessa multiplicidade. A teoria da linhagem germinal propunha a existência de um gene distinto para cada cadeia diferente de anticorpo e esse repertório seria transmitido por herança para as gerações subseqüentes. Em

---

<sup>9</sup>Regiões Fab.

<sup>10</sup>IgM e IgE possuem uma quarta região  $C_H^4$ .

contrapartida, as teorias da diversificação somática acreditavam na existência de um número limitado de genes codificadores das regiões V que seriam herdadas e a diversidade decorreria de mutações desses genes nas células B.

Hoje se sabe que essas teorias estão ambas parcialmente corretas. O repertório de anticorpos é gerado durante o desenvolvimento das células B por rearranjos do DNA. Estes rearranjos combinam diferentes segmentos gênicos da região V [38]. Outro fator que potencializa essa diversidade são as hipermutações somáticas que ocorrem nos linfócitos B ativados.

### 2.11.1 A Diversidade na Região V

A região V é codificada pelo rearranjo de mais de um segmento gênico. Dentro da cadeia leve, a porção variável requer dois segmentos distintos de DNA para a sua codificação. O primeiro segmento, denominado segmento gênico V codifica a maior parte do domínio variável, que compreende os primeiros 95-101 aminoácidos da cadeia leve. O segundo segmento codifica a região de junção, que corresponde ao restante do domínio V constituído de até 13 aminoácidos. Esse último segmento é denominado junção ou segmento gênico J [38]. A junção desses dois segmentos forma um fragmento de DNA capaz de codificar todo o domínio variável da cadeia leve.

Em uma célula B não ativada, as seqüências de DNA codificadoras da região V ficam espacialmente separadas por uma distância considerável daqueles genes codificadores da região C. Quando os linfócitos B amadurecem, ocorre um fenômeno conhecido como recombinação somática, que rearranja os genes das imunoglobulinas fazendo com que eles fiquem próximos. Antes do rearranjo, os segmentos J se localizam próximos aos segmentos gênicos da região C. A união dos segmentos V aos J aproxima V e C. O segmento J de uma região V rearranjada fica então separado da região C por apenas um pedaço de DNA não codificante (*íntron*). O RNA mensageiro da cadeia leve completa é obtido através do processamento do RNA transcrito. Este processamento une os genes codificadores das porções constantes e variáveis da cadeia leve. O procedimento completo descrito, desde o rearranjo até a formação da cadeia peptídica final pode ser visto na figura 2.15, quadro à esquerda.

A cadeia pesada é codificada por três segmentos gênicos  $V_H$ ,  $J_H$  e  $D_H$ <sup>11</sup>. O segmento novo  $D_H$ , conhecido como segmento gênico de diversidade fica localizado entre os segmentos  $V_H$  e  $J_H$ . A recombinação somática que produz regiões V da cadeia pesada ocorre em duas etapas. Na primeira,  $D_H$  é reunido a  $J_H$  seguido pelo rearranjo que une  $V_H$  ao combinado  $D_HJ_H$ . Novamente, o processamento do RNA junta as seqüências das regiões V e C, como pode ser visto na figura 2.15, quadro à direita.

<sup>11</sup>A designação dos segmentos V, D e J vem acompanhada da letra H - do inglês *heavy* - para deixar explícita a distinção entre os segmentos codificadores da cadeia leve e pesada.

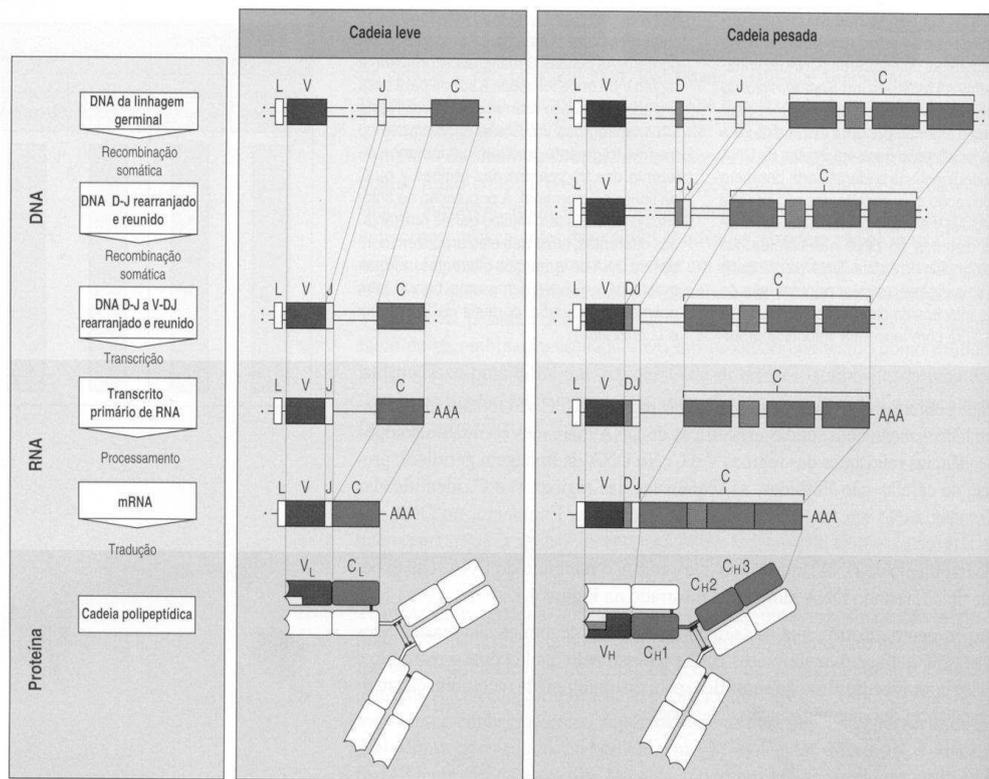


Figura 2.15: A geração da diversidade das regiões V do anticorpo (adaptado de [38]).

### Os Segmentos Gênicos da Região V

Existem múltiplas variações dentro de cada segmento gênico V, D, J e C, o que proporciona uma maior diversidade na seleção aleatória de cada integrante codificador das regiões V da imunoglobulina. Os segmentos gênicos são organizados em três grupos. São os genes de cadeias leves  $\kappa$  e  $\lambda$  e de cadeias pesadas H. Os grupos possuem organizações internas diferentes e se localizam em cromossomos distintos, como mostrado no quadro da figura 2.16. A peculiaridade de cada grupo com relação à ordem interna está relacionada à disposição dos segmentos V, D, J, e C. A distribuição dos segmentos dentro de cada grupo está ilustrada na figura 2.17. É possível observar na mesma figura que há pedaços de genes entre os segmentos codificadores da região V. Estes fragmentos representam segmentos não funcionais, conhecidos como pseudogenes e cuja recombinação gera codificadores inválidos.

Há uma divisão dos segmentos gênicos V humanos em famílias. Membros da mesma família devem possuir obrigatoriamente um mínimo de 80% de semelhança entre si. Os segmentos  $V_H$  e  $V_\kappa$  pertencem a sete famílias distintas e os segmentos  $V_\lambda$  a oito. Dentro de cada família pode ainda haver uma divisão em clãs. Nos humanos, os segmentos  $V_H$  se inserem em três clãs.

Número de segmentos gênicos funcionais nos <i>loci</i> de imunoglobina humana			
Segmento	Cadeias leves		Cadeia pesada
	$\kappa$	$\lambda$	H
Variável (V)	40	30	65
Diversidade (D)	0	0	27
Junção (J)	5	4	6

Figura 2.16: Organização dos segmentos gênicos de cadeias leve e pesada dentro dos grupos  $\lambda$  (cromossomo 22),  $\kappa$  (cromossomo 2) e H (cromossomo 14) em humanos (adaptado de [38]).

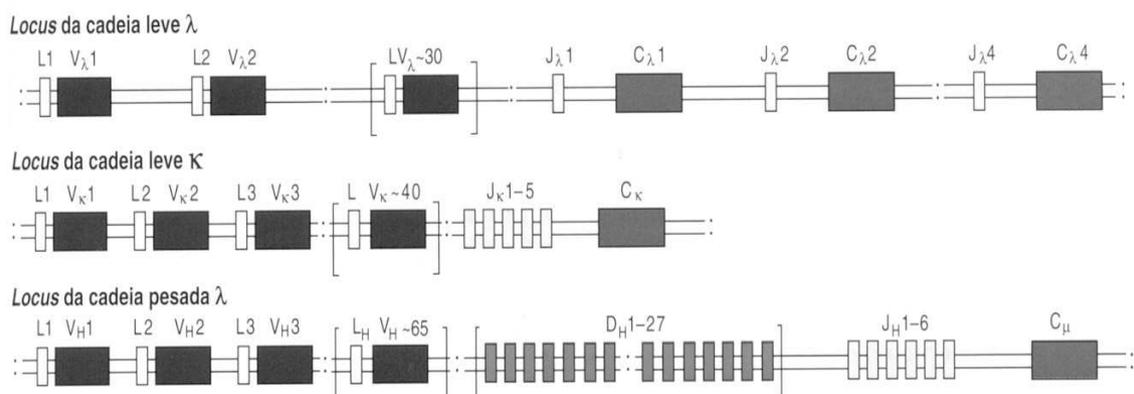


Figura 2.17: Disposição dos segmentos gênicos de cadeias leves e pesadas em humanos (adaptado de [38]).

### **2.11.2 Processos Geradores da Diversidade dos Anticorpos**

Há quatro mecanismos de geração da diversidade dos anticorpos. O primeiro já foi discutido anteriormente e decorre do processo de recombinação somática. Envolve a multiplicidade de segmentos gênicos da região V e o número grande de combinações que podem ser utilizadas nos rearranjos. O segundo mecanismo, denominado diversidade juncional, é consequência da adição e subtração de nucleotídeos nas junções entre os diferentes segmentos gênicos durante o processo de recombinação. Em seguida têm-se as muitas possibilidades de combinação entre as cadeias leve e pesada para obtenção da molécula completa do anticorpo. Finalmente, já nos linfócitos B maduros, ocorre a hipermutação somática, que introduz mutações nos genes de regiões V rearranjados [38].

### **2.11.3 Hipermutações Somáticas**

As hipermutações, além de serem um mecanismo gerador de diversidade na população de anticorpos, são responsáveis também pelo refinamento da afinidade do anticorpo com relação ao seu antígeno específico. Elas operam sobre células B maduras localizadas nos órgãos linfóides secundários. Sua função é introduzir mutações pontuais nas regiões V dos genes rearranjados que codificam as cadeias leves e pesadas, em uma taxa muito elevada [38].

O objetivo desse mecanismo é produzir imunoglobulinas mutantes na superfície das células B, a fim de que algumas dessas variações resultem em anticorpos de maior conformidade físico-química contra o antígeno quando comparados à imunoglobulina original. Aqueles linfócitos capazes de produzir imunoglobulinas com poder de neutralização superior tendem a ser privilegiados no processo de escolha das células B que irão se diferenciar em plasmócitos. Este fenômeno de seleção dos melhores anticorpos é chamado maturação de afinidade.

#### **Mecanismo Regulador das Hipermutações Somáticas**

As hipermutações somáticas ocorrem nos centros germinativos, que são locais que proporcionam um ambiente facilitador da maturação da afinidade das células B. Supõe-se que as hipermutações ocorram aleatoriamente ao longo de toda a região V, apesar de haver certos pontos com maior tendência às mutações. Dado esse caráter aleatório, há a possibilidade de ocorrer modificação em alguns dos 600-700 pares de bases disponíveis nas regiões V. A probabilidade de haver hipermutação é de  $1 \times 10^{-3}$  por par-base das regiões variáveis, o que equivale a dizer que, em média, é introduzida uma mutação a cada divisão celular. Ocorre que, a maioria das alterações produzidas nesses genes das células B geraria imunoglobulinas de baixa-afinidade, inválidas ou auto-reativas. A questão levantada em [53] está em entender como um número relativamente limitado de hipermu-

tações produtoras de alta afinidade pode ainda assim sobrevir com tanta frequência dentro do organismo.

Para responder a esta questão, Perelson e Oprea [53] desenvolveram um modelo de expansão das células B e mutação somática que mostra que a reciclagem frequente de centrócitos <sup>12</sup> antígeno-selecionados em centroblastos <sup>13</sup> pode levar a uma maturação de afinidade eficiente. O modelo sugere uma reação no centro germinativo na qual a saída das células é reduzida a uma dissociação do centro germinativo, seguida pela liberação de centrócitos na sua periferia. Isto seria vantajoso para a geração de memória de alta afinidade. Outros modelos que tentam explicar o sucesso do processo de maturação da afinidade podem ser vistos em [67, 55].

Um outro ponto ainda não esclarecido está em como preservar as boas combinações de genes havendo uma probabilidade de mutação tão alta. Uma célula portadora de combinação de alta afinidade antigênica, ao longo de sua reprodução, sob a mesma taxa de mutação, pode ter sua afinidade destruída pelo acúmulo de novas mutações indesejadas. Dessa forma, acredita-se que o processo de hipermutação somática seja constituído por um curto pico de mutação, seguido de um período de intervalo para a seleção e expansão clonal. A seleção deve regular o processo de hipermutação para preservar as células de alta afinidade e permitir que células de baixa afinidade permaneçam sendo mutadas [42, 43].

## 2.12 A Geração de Diversidade dos Receptores das Células T (TCR)

Como pode ser visto na figura 2.18, a geração da diversidade dos TCR é muito semelhante à das imunoglobulinas. As cadeias  $\alpha$  contêm segmentos gênicos  $V_\alpha$  e  $J_\alpha$ . Os genes da cadeia  $\beta$  têm segmentos  $V_\beta$ ,  $D_\beta$  e  $J_\beta$ . O rearranjo dos genes que codificam os receptores dos linfócitos T ocorre no timo.

A principal diferença entre os genes codificadores dos TCR e das imunoglobulinas está na complexidade dos genes das regiões constantes. As imunoglobulinas, além de se ligarem a antígenos, ativam mecanismos efetores através dos isótipos das regiões constantes. Como os TCR não medeiam diretamente as funções efectoras das células T, os genes codificadores das regiões constantes são mais simples do que aqueles das imunoglobulinas.

---

<sup>12</sup>Centrócitos são pequenas células B, não proliferantes nos centros germinativos, que derivam dos centroblastos. Eles amadurecem em células plasmáticas formadoras de anticorpos, células B de memória ou ainda podem sofrer apoptose, dependendo da inter-relação de seu receptor com o antígeno [38].

<sup>13</sup>Centroblastos são células grandes, em rápida divisão, encontradas nos centros germinativos, nas quais se acredita ocorrer a hipermutação somática. Os plasmócitos e as células B de memória derivam dessas células [38].

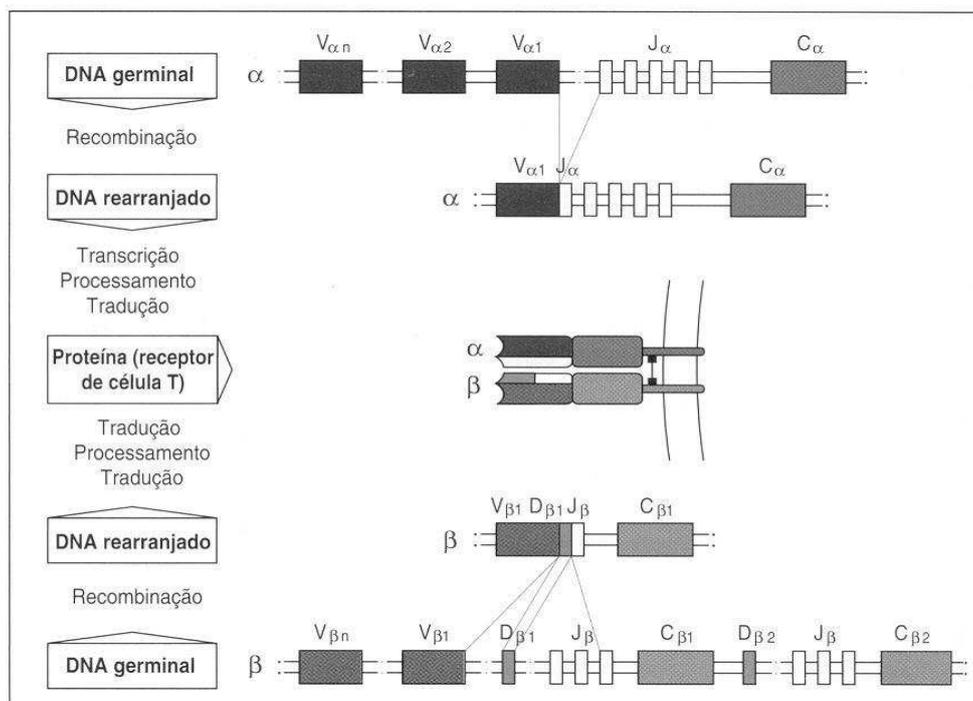


Figura 2.18: Geração da diversidade dos TCR (adaptado de [38]).

## 2.13 Distinção Próprio/Não Próprio

O repertório de linfócitos presente no organismo é amplo o suficiente para reconhecer todos os possíveis invasores que apareçam ao longo de sua existência, o que é próprio do organismo *self* e o que é próprio e apresenta anomalias.

A especificidade dos linfócitos é determinada nas primeiras etapas de sua diferenciação, quando o RNA mensageiro codificador dos receptores das células B e T é montado a partir de segmentos gênicos rearranjados.

Dado potencial criador dos rearranjos gênicos, associado aos outros fatores que geram diversidade, é possível admitir três classes de repertórios celulares [16]. O *repertório potencial* representa tudo que pode ser gerado pelos mecanismos de expressão genética associados às variações obtidas pelas mutações. O *repertório expresso* engloba moléculas disponíveis para tornarem-se receptores e o *repertório ativo* são os linfócitos que participam nas respostas imunes.

Em uma fase seguinte à expressão do receptor antigênico na superfície do linfócito, há uma avaliação que verifica as propriedades de reconhecimento antigênico da célula com relação às moléculas do ambiente. Os resultados desse teste de especificidade e afinidade do receptor determinam se o linfócito está apto a sobreviver ou se deve ser eliminado.

### **2.13.1 Seleção Positiva e Seleção Negativa**

Em geral, linfócitos em desenvolvimento que interagem fracamente com antígenos próprios, ou auto-antígenos, recebem sinais de sobrevivência. A escolha dos linfócitos que permanecerão vivos é chamada seleção positiva. Particularmente para as células T, a seleção positiva escolhe aquelas cujos receptores são capazes de reconhecer e se ligar a moléculas de MHC-próprio [16, 17], o que constitui uma premissa para que uma célula T seja capaz de armar uma resposta contra um antígeno. A seleção positiva nas células B maduras ocorre após a hipermutação, quando a célula filha mutante é mais eficiente na neutralização do anticorpo e é preferida para expansão. Produtos ineficientes da hipermutação sofrem apoptose, que é a morte celular programada.

Linfócitos de alta reatividade com antígenos próprios são sinalizados para morte, o que caracteriza a seleção negativa. Desta forma, linfócitos auto-agressivos são removidos do repertório antes de amadurecerem e desencadearem reações auto-imunes. Esta seleção estabelece a tolerância imunológica para auto-antígenos.

A sobrevivência dos linfócitos para a fase de ativação é condicionada a sinais enviados para os seus receptores. Aqueles que não receberem esses sinais são programados para a morte. A maior parte das combinações dos genes codificadores gera linfócitos auto-reativos. Desta forma, somente uma pequena parcela de linfócitos estão destinados a amadurecer e compor o repertório linfocitário do sistema adaptativo. Ainda assim, esta pequena fração restante é capaz de responder, em termos práticos, a qualquer antígeno apresentado [38].

## **2.14 A Teoria da Seleção Clonal**

A teoria da Seleção Clonal [7, 2] conjecturou que o antígeno, ao penetrar no corpo do hospedeiro, recrutaria um subconjunto de linfócitos capazes de reconhecê-lo. Linfócitos com maior reatividade com o antígeno seriam selecionados para expansão em clones em detrimento daqueles com menor capacidade de neutralização. Cada clone seria portador de um único tipo de receptor que não estaria presente em nenhum outro clone.

Uma das contribuições mais importantes dessa teoria, inicialmente proposta por F. M. Burnet [7], foi a profetização, em 1957, de que clones auto-reativos deveriam ser destruídos durante a seleção [62]. Esta foi a primeira vez que o paradigma da distinção entre o que é próprio e o que é estranho foi mencionado na Imunologia. Em [78], há um ensaio muito interessante contextualizando a teoria burnetiana hoje, seus erros e acertos, o porquê de ainda ser adotada e como seria uma nova teoria para substituí-la. Partindo da idéia da deleção de clones de células que agrediriam o próprio organismo, Burnet bloqueou qualquer possibilidade de reações auto-imunes e interações dos linfócitos entre si e com elementos do que é próprio.

Hoje em dia, sabe-se que há doenças auto-imunes e que os linfócitos B e T comunicam entre si. Os linfócitos T precisam do MHC para reconhecerem o antígeno. Por isso, são auto-reativos. A questão levantada em [78] foi como, mesmo com tantos exemplos contrários o modelo se mantém firme, orientando a pesquisa e o ensino da imunologia. A resposta dada foi que essa teoria inseriu a Imunologia no Neodarwinismo, sugerindo os princípios de seleção natural e casualidade [12]. Desta forma, rejeitar a teoria de Burnet seria negar a teoria evolutiva dominante na biologia. Vaz [78] acredita na sedimentação de uma teoria mais moderna como consequência de novas descobertas envolvendo a organização global do SI auxiliadas por pesquisas interdisciplinares.

Portanto, o quadro atual postula que, em termos gerais, a seleção clonal se inicia quando o agente estranho penetra o organismo e ativa o complexo inato. Aqueles linfócitos que apresentam maior poder de ligação com os determinantes antigênicos se proliferam nos centros germinativos, na chamada expansão clonal. Ambas células B e T passam pelo processo de seleção clonal, com a diferença que linfócitos do tipo B sofrem hipermutações somáticas durante a expansão clonal. Aquelas células ou anticorpos que apresentarem melhor ligação ao antígeno, sem serem auto-reativas, recebem estímulos para sobrevivência e proliferação, em detrimento de linfócitos menos aptos, que sofrem apoptose.

Um ponto importante a ressaltar é que a Imunologia não mais assume a seleção clonal como uma teoria. Ela de fato acontece na porção adaptativa do sistema imune e seu comportamento já foi largamente estudado e comprovado. Não há também, em livros modernos, como em [38], o assunto “seleção clonal” como um tópico. Toda a fenomenologia descrita acima encontra-se “diluída” no texto. Esta separação é feita, entre outros motivos, para destacar a semelhança com o Darwinismo.

## 2.15 A Arquitetura do SI e o Sistema Global de Defesa

A união de todos os componentes descritos nas seções anteriores constitui o sistema global de defesa dos mamíferos contra patógenos. Observando o SI já como um todo, é possível distinguir em sua arquitetura um esquema de múltiplas camadas contendo mecanismos defensivos e reguladores espalhados por todo seu domínio [16].

A camada que constitui a primeira linha global de defesa do organismo, as *barreiras físicas*, é formada por órgãos e sistemas que terão o primeiro contato com o antígeno. Sua função é impedir que o antígeno penetre o interior do organismo. É composta pela pele, que trabalha como um escudo protetor contra qualquer invasor; pelo sistema respiratório, onde há apreensão de partículas através dos pêlos, mucosas nasais e ativação de mecanismos de tosse e espirro; e pelas mucosas do trato digestivo.

Em seguida, é possível identificar *barreiras bioquímicas*, como a saliva, suco gástrico, lágrima, pH e temperatura corporais que ajudam a eliminar e proporcionar um ambiente

desfavorável para o estabelecimento dos antígenos no hospedeiro. As duas camadas seguintes são representadas pelas partes inata e adaptativa do SI.

As APCs do sistema inato percorrem o organismo a procura de patógenos que porventura tenham conseguido superar barreiras físicas e bioquímicas do organismo. Quando um antígeno é detectado, a APC o fagocita e digere em peptídeos. Parte desses peptídeos se liga a proteínas do MHC, formando um complexo MHC/peptídeo que se dirige para a superfície da membrana da APC. A APC migra para os órgãos linfóides secundários e apresenta esse complexo MHC/peptídeo aos linfócitos T. Aqueles linfócitos cujos receptores forem capazes de reconhecer a combinação exibida se tornam ativados, proliferam e secretam linfocinas capazes de mobilizarem outros elementos do sistema imune, dentre os quais estão os linfócitos B. As células B ativadas se expandem em clones de plasmócitos que secretarão anticorpos contra o antígeno. Os anticorpos secretados têm tanto a função de neutralizar o antígeno como ativar mecanismos efetores responsáveis pela eliminação desse agente agressor. Algumas células B e T, durante a sua expansão, transformam-se em células de memória capazes de potencializar a reação imune frente a uma segunda invasão desse patógeno. Em certos casos, os linfócitos B são capazes de reconhecer o antígeno sem a necessidade das linfocinas liberadas pelas células T.

Os conceitos vistos neste capítulo inspiraram o desenvolvimento de técnicas inteligentes para a resolução de certos problemas computáveis. Em particular, a área de sistemas imunológicos artificiais, que será abordada no próximo capítulo.

# Capítulo 3

## Sistemas Imunológicos Artificiais

*Este capítulo apresenta motivações, conceitos e paradigmas adotados para o desenvolvimento da Engenharia Imunológica.*

### 3.1 Introdução

Comparada a outros paradigmas de inteligência computacional, a área do conhecimento envolvendo Sistemas Imunológicos Artificiais (SIAs) é ainda muito nova e pouco explorada. Entretanto, ela apresenta uma quantidade vasta de conceitos explorados na elaboração de soluções computacionais para os mais diversos fins. Esta abrangência fez com que se formassem diversas escolas, cada uma envolvendo um subconjunto de paradigmas dos SIs para aplicações distintas. Estabelecer, portanto, um algoritmo ou conjunto de algoritmos que definam a base do que seria um SIA é uma tarefa bastante difícil, pois cada descoberta a respeito dos mecanismos imunitários biológicos pode ser fonte de inspiração para a elaboração de novos tipos de soluções.

Em [16] e [17], uma tentativa mais abrangente de se estabelecer um “esqueleto” para o projeto de um SIA foi proposta. Nelas, os autores compilaram um conjunto de algoritmos envolvendo imunologia teórica e SIAs, introduzindo o termo engenharia imunológica (EI). A EI foi definida como “*processo de meta-síntese utilizado para extrair idéias do sistema imunológico a fim de construir ferramentas computacionais para resolução de problemas*”. Estas ferramentas se propõe a auxiliar tanto no desenvolvimento da Imunologia quanto da Engenharia.

O avanço da Imunologia é proporcionado através de modelos computacionais que simulam aspectos do sistema imunitário [17], efeitos de vacinas, conformação de moléculas de fármacos, doenças infecciosas, bibliotecas genéticas [36, 54, 60, 24], entre outros.

Pesquisas nas áreas de engenharia procuram desenvolver soluções inspiradas nos mecanismos imunes para, por exemplo, realizar reconhecimento de padrões, descoberta de conhecimento em bases de dados, busca e otimização, robótica, controle, segurança,

aprendizado de máquina, vida artificial, diagnósticos de erros em equipamentos e planejamento.

As seções seguintes apresentam alguns dos principais modelos artificiais do SI e algoritmos relevantes desenvolvidos na área até o presente momento.

## 3.2 Modelos do Sistema Imunológico

### 3.2.1 A Teoria das Redes Idiotípicas

Em sua teoria, Jerne [39] partiu da premissa de que dentro do sistema imune de um indivíduo, qualquer idiótopo dos anticorpos pode ser reconhecido por um conjunto de parátomos, e qualquer parátomo pode reconhecer um conjunto de idiótopos. Havendo repertórios suficientemente grandes de parátomos e idiótopos, o sistema imune pode ser visto como uma rede. Haverá sempre parátomos que reconhecem grupos de idiótopos, e idiótopos reconhecidos por conjuntos de parátomos. Quando um antígeno penetra no organismo do indivíduo, ele passa a fazer parte da rede.

### 3.2.2 O Modelo *Shape Space*

Este modelo foi concebido por Perelson e Oster [57], a fim de descrever as interações entre anticorpos do SI e antígenos. Mais tarde, a idéia foi adaptada para o uso em SIAs. Partindo das características bioquímicas que o anticorpo e antígeno devem ter para se dar a neutralização, como conformação molecular, força de ligação, área de complementaridade e distribuição das cargas, pode-se determinar um formato geral para as moléculas.

Assume-se que é possível descrever o formato do anticorpo por um conjunto de  $L$  parâmetros [17, 58]. Portanto, um espaço  $L$ -dimensional  $S$ , chamado *shape space*, determina a região ligante do anticorpo ou receptor linfocitário de acordo com suas propriedades de adesão. Por consequência, o mesmo *shape space* pode ser usado para descrever os determinantes antigênicos, ou o próprio antígeno, se ele for do mesmo tamanho do anticorpo.

Se um sistema tem um repertório  $N$  de anticorpos, então  $S$  tem  $N$  pontos. Cada ponto encerra um volume  $V_\epsilon$  no espaço sobre o qual ele tem poder de ligação. Antígenos cujo determinante antigênico se localiza em  $V_{\epsilon i}$  são neutralizados pelo  $i$ -ésimo anticorpo. Um esquema de como ficaria o espaço  $S$  pode ser visto na Figura 3.1.

Matematicamente, anticorpos e determinantes antigênicos poderiam ser representados por um vetor  $X$  de  $n$  dimensões, onde cada dimensão determina um atributo do problema. Há quatro principais *shape spaces* em SIAs: de atributos reais, inteiros, *Hamming* e simbólico [17].

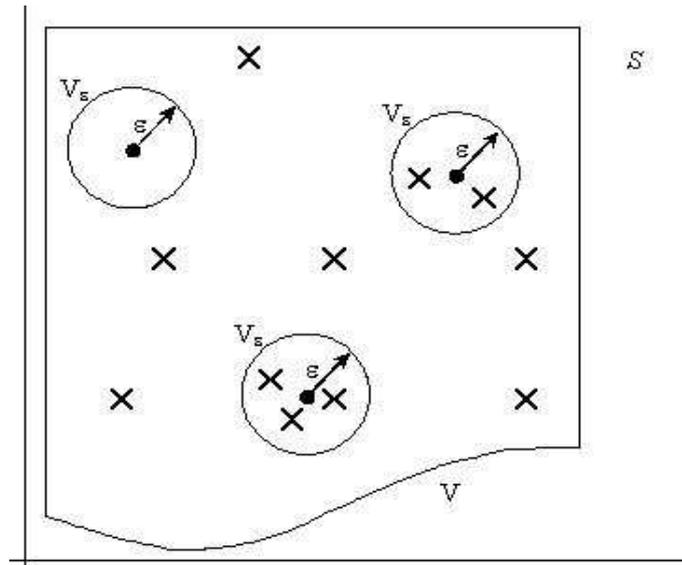


Figura 3.1: Esquema de um *Shape space*, onde  $\bullet$  representam os anticorpos e  $\times$  os antígenos. Os círculos que envolvem os anticorpos são a sua área de atuação. Qualquer antígeno presente em uma dessas áreas terá um anticorpo correspondente que irá neutralizá-lo. Quanto mais próximo o antígeno estiver do anticorpo, maior será a força de ligação.

### 3.2.3 Modelos Evolucionários

Em [26], Forrest e Perelson propuseram um modelo de SIA utilizando antígenos e anticorpos representados por cadeias binárias. O modelo tinha como objetivo estudar o reconhecimento de padrões em indivíduos e espécie. Para isto, anticorpos e antígenos eram inicializados aleatoriamente. Um algoritmo genético (AG) ficaria responsável por determinar quais anticorpos seriam úteis ao SIA. A avaliação dos anticorpos era dada segundo uma função que determina a sua capacidade de ligação contra os antígenos. Identificando populações de antígenos contendo segmentos binários em comum, a meta do modelo era obter um anticorpo capaz de reconhecer padrões existentes na população de patógenos.

Um aprimoramento deste modelo, descrito em [36] por Forrest, Perelson e Hightower, introduziu o conceito de bibliotecas genéticas para gerar anticorpos. Foram utilizadas quatro bibliotecas compostas por oito elementos de 16 bits. Um segmento de cada biblioteca era escolhido aleatoriamente, e sua junção com os outros três determinava o anticorpo. A aptidão do anticorpo foi definida como sua capacidade geral de reconhecer antígenos. Outro trabalho envolvendo bibliotecas genéticas foi desenvolvido em [24] para simular a filogênese de uma espécie artificial.

Em um quarto trabalho [54] de Oprea e Forrest, cada indivíduo era constituído de uma biblioteca de  $A$  anticorpos de tamanho  $L$  bits. Os antígenos possuíam igualmente cromossomos de tamanho  $L$ . Para cada patógeno, o anticorpo responsável pela sua neutralização era aquele com maior energia de ligação. No AG utilizado, foi construída uma população inicial aleatória de 50 bibliotecas de mesmo tamanho. Para produzir uma geração seguinte, a seleção era feita por *rank* e o esquema de substituição geracional. A

avaliação do indivíduo (biblioteca) era dada pela energia de ligação entre os patógenos e os anticorpos. Esta energia foi definida por uma curva de Gauss. O melhor indivíduo seria a biblioteca que produzisse anticorpos com maior força de ligação.

## 3.3 Algoritmos e Processos

### 3.3.1 Algoritmo CLONALG

O algoritmo CLONALG [16] e [17] parte do princípio de que quando um linfócito reconhece um antígeno com um certo grau de afinidade (dada por uma medida de distância – Euclidiana, por exemplo), ele tende a proliferar e gerar clones. Para células B, ocorrem hipermutações que realizam a maturação da afinidade, o que faz com que o reconhecimento seja cada vez mais refinado. Linfócitos B selecionados para reprodução (aquelas com maior afinidade) são mantidos como células de memória. O pseudo-código do modelo pode ser visto em Algoritmo 1.

---

#### Algoritmo 1 *CLONALG*

---

- ◊ Inicializar aleatoriamente uma população de células P.
  - ◊ Para cada antígeno faça:  
**laço**
    - ◊ Apresentar o antígeno para a população P e determinar a afinidade com relação a cada linfócito;
    - ◊ Gere um número  $N_c$  de clones para as células com maior afinidade, sendo que quanto maior a afinidade, maior o número de clones;
    - ◊ Mute cada clone proporcionalmente à aptidão da célula: quanto maior a afinidade, menor a taxa de mutação e vice-versa;
    - ◊ Adicione as células modificadas a P e selecione as melhores para formarem a memória celular;
    - ◊ Para cada clone, selecione a célula com a maior aptidão e calcule a aptidão média da população selecionada. Determine a aptidão de todos os indivíduos da população;
    - ◊ Substituir células de baixa afinidade por outras geradas aleatoriamente;
    - ◊ Definir um critério de parada;
  - fim laço**
- 

### 3.3.2 Algoritmo *OPT-AINET*

*aiNet* [18] é um modelo de rede imune artificial originalmente desenvolvido para fazer compressão de dados. Uma versão do *aiNet* foi desenvolvida para resolução de problemas de otimização, o *opt-aiNet* [71, 15]. Dentro deste novo paradigma, é assumida a seguinte terminologia:

**Célula** de Rede: é o indivíduo da população. Cada célula é um vetor de valores reais no espaço Euclidiano.

**Afinidade:** é a distância Euclidiana entre duas células.

**Clone:** células filhas que são cópias idênticas da célula genitora. Estas células se tornam variações de seus pais por meio de mutações somáticas.

---

**Algoritmo 2** *opt-aiNet*

---

◇ Inicializar aleatoriamente uma população de células. O número inicial de células não é relevante;

◇ Enquanto o critério de parada não for atingido, faça:

**laço**

◇ Determine a aptidão de cada célula e normalize o vetor de aptidão;

◇ Gere um número  $N_c$  de clones para cada célula de rede;

◇ Mute cada clone na proporção inversa à aptidão da sua célula pai, mantendo a pai;

◇ Para cada clone, selecione a célula com a maior aptidão e calcule a aptidão média da população selecionada. Determine a aptidão de todos os indivíduos da população;

◇ Se o erro médio da população não for muito diferente daquele da primeira iteração, então continue. Senão, retorne ao início do laço;

◇ Determine a afinidade de todas as células na rede. Suprima todas menos a de maior aptidão dentre as células cujas afinidades são menores que o *threshold* e determine o número de células de memória;

◇ Introduza uma porcentagem  $d\%$  de células geradas aleatoriamente;

**fim laço**

---

O *opt-aiNet* pode ser resumido segundo o Algoritmo 2. Há algumas características interessantes do *opt-aiNet* que podem ser destacadas:

- apresenta um mecanismo de seleção elitista;
- a cardinalidade da população é determinada pela supressão e introdução de diversidade;
- quanto maior for o número de ótimos que ele localizar maior a sua capacidade de achar mais ótimos;
- apresenta custo computacional baixo.

O critério de parada é baseado no tamanho da população de células de memória. Se ela não se altera ao longo das últimas rodadas, então o equilíbrio foi atingido. Pode-se considerar que as células de memória remanescentes sejam as soluções para o problema.

Há vários outros algoritmos imunes baseados na medula óssea, timo, redes idiotípicas, entre outros.

### 3.4 SI e Complexidade

Segundo a teoria de sistemas, pode-se definir um sistema complexo como aquele onde seus indivíduos coletivamente apresentam características e comportamentos diferentes dos apresentados quando somente um integrante do sistema é analisado individualmente. Isto significa que a dinâmica de um sistema complexo emerge das diversas interações entre as suas partes e do papel que cada uma desempenha para uma resposta coletiva e manutenção da estabilidade do conjunto. O Sistema Imune é um exemplo de sistema complexo, onde, observando cada um de seus integrantes, é impossível ter idéia da dimensão da sua atuação dentro do organismo. A ação cooperativa dos subsistemas inato e adaptativo é o que lhe confere a capacidade de proteger um indivíduo das agressões externas.

O interessante de analisar o SI sob o ponto de vista sistêmico está na possibilidade de enquadrá-lo em categorias e identificar semelhanças a sistemas já analisados, na tentativa de extrair mais informação sobre o seu funcionamento. Sabe-se que, pelo fato do SI ser capaz de interagir com outros sistemas dentro do organismo e com agentes invasores, sob o ponto de vista das fronteiras, ele pode ser classificado como um sistema aberto.

A reinfestação do organismo por um antígeno acarreta uma segunda resposta imunológica mais rápida e eficiente contra esse patógeno. A filtração da linfa ao passar pelos linfonodos tem sua eficácia condicionada ao número de macrófagos presentes no momento, fato que determina respostas imunes com intensidades diferentes, dependendo da disposição e estado geral dos integrantes do SI. Os linfócitos, de uma forma genérica, são antígeno-específicos e estão presentes em quase todos os sistemas do organismo. Estes são exemplos que conferem ao SI um caráter não-linear, ou seja, a resposta de cada elemento do sistema nem sempre é proporcional ao estímulo que o mesmo recebe do meio ou de outro elemento. O fato das respostas imunes não partirem ou serem controladas por um elemento central destaca no SI a qualidade de ser espacialmente distribuído.

O SI também é classificado como dinâmico, probabilístico e auto-organizável. É dinâmico porque evolui com o tempo e tem a capacidade de muitas vezes se adaptar a novas circunstâncias. A questão da diversidade dos anticorpos e dos receptores das células T, juntamente com as mutações somáticas, são responsáveis pelo seu caráter aleatório.

Auto-organização é o processo no qual a ordem do sistema aumenta gradativamente por meio da ajuda mútua entre seus elementos. Esta cooperação surge da necessidade de uma resposta a estímulos externos ou para satisfazer propriedades naturais de seus elementos. Muitas vezes, para que se estabeleça uma ordem satisfatória do sistema dentro de um certo contexto, é necessário que algumas partes do sistema não funcionem, ou percam parcialmente suas funções, para que outras possam desempenhar seu papel mais adequadamente. Nota-se, portanto, que é possível haver no sistema estágios de cooperação e competição entre os indivíduos, objetivando a evolução do sistema global.

No próximo capítulo serão vistos alguns conceitos sobre classificação de dados necessários para o entendimento das propostas apresentadas neste trabalho.

# Capítulo 4

## Classificação

*Este capítulo tem como objetivo apresentar um resumo sobre alguns tópicos relevantes dentro da pesquisa atual envolvendo classificação.*

### 4.1 Introdução

Classificar, dentro do contexto da descoberta do conhecimento, significa analisar um conjunto de atributos que caracterizam um determinado objeto e, a partir daí, determinar-lhe um rótulo. Este processo é dividido em duas partes.

A primeira parte, o treinamento, consiste na construção de um modelo capaz de descrever as principais características de uma classe, a partir das informações obtidas pelos atributos contidos nas tuplas ou registros de uma tabela em uma base de dados. As tuplas contêm, ainda, um rótulo que determina a categoria ou classe a qual o registro pertence [34].

Os exemplos selecionados na base de dados para o aprendizado do modelo são o conjunto de treinamento. A seleção destes dados é feita aleatoriamente, e como o classificador é construído utilizando as informações contidas sobre as classes no campo rótulo, esse aprendizado é supervisionado.

Realizada a primeira parte, a que se segue é de fato a classificação. Nesta etapa, fornecidos modelo e novas amostras de dados, objetiva-se obter o maior valor de acurácia na determinação das classes dos exemplos que não foram apresentados ao classificador, durante o treinamento.

Nas seções seguintes são estudadas as principais questões relacionadas a desbalanceamento de classes, algumas abordagens de soluções e estratégias para aumento e redução do conjunto de treino de acordo com o tipo de problema a ser resolvido.

## 4.2 O Problema do Desbalanceamento de Classes

Sabe-se que bases de dados que envolvem problemas reais geralmente apresentam quantidades de exemplos em suas classes em proporções diferentes. Muitas vezes, as quantidades de registros em cada classe podem comprometer o aprendizado, fazendo com que o método classificador não seja eficiente na predição de uma certa categoria. Estes casos que aparecem pouco na base de dados podem ser ruídos ou objetos raros.

Objetos raros em problemas de mineração de dados são de grande valor, quando, por exemplo, se deseja detectar alguma anomalia [13], como fraudes fiscais, problemas em transmissões de imagens de satélites, invasões a servidores de redes, peças de motores defeituosas, entre outros [87].

Há dois tipos de raridades no contexto do estudo de mineração de dados, segundo Weiss [87]. O primeiro é caracterizado por classes raras, que constituem o problema de desbalanceamento de classes. O segundo tipo são os casos raros dentro de uma classe, ou seja, registros que representam uma região pequena, porém importante, do espaço de instâncias.

Minerar casos e classes raras pode ser difícil por vários motivos, como a utilização de métricas impróprias, a pouca ou nenhuma ocorrência de exemplos de uma determinada classe para serem apresentados para o classificador, fragmentação de dados, ruídos e um viés indutivo (*inductive bias*) inapropriado.

### 4.2.1 Métricas de Avaliação de Classificadores

Para problemas de duas classes, as métricas são baseadas na matriz de confusão, uma ferramenta que informa os tipos de erros e acertos de um classificador [22]. As classes são ditas positivas e negativas e a matriz de confusão possui quatro valores computados em termos dos valores preditos, que são:

- TP (*true positives* - positivos verdadeiros): quantidade de elementos positivos classificados como positivos;
- TN (*true negatives* - negativos verdadeiros): quantidade de elementos negativos classificados como negativos;
- FP (*false positives* - positivos falsos): quantidade de elementos negativos classificados como positivos;
- FN (*false negatives* - negativos falsos): quantidade de elementos positivos classificados como negativos;

Outras métricas mais comuns utilizadas para determinar a eficiência de classificador proposto são:

- Acurácia (*acc*): taxa de decisões corretas feitas pelo classificador:

$$acc(val) = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

- Sensibilidade (*sens*): mede o quanto um classificador consegue reconhecer exemplos positivos:

$$sens = \frac{TP}{TP + FN} \quad (4.2)$$

- Especificidade (*spec*): mede o quanto um classificador consegue reconhecer exemplos negativos:

$$spec = \frac{TN}{TN + FP} \quad (4.3)$$

- Acurácia Balanceada (*accB*): Indica o valor de acurácia levando em consideração o desbalanceamento de classes:

$$accB = \frac{(sens + spec)}{2} \quad (4.4)$$

- Precisão (*prec*): é a taxa de exemplos positivos classificados que realmente são positivos:

$$prec = \frac{TP}{TP + FP} \quad (4.5)$$

- *F-measure* (*FMea*): é a média harmônica da sensibilidade e precisão. Neste estudo, o parâmetro  $\beta$  é igual a zero, o que significa que sensibilidade e precisão têm a mesma importância:

$$FMea = \frac{(\beta^2 + 1) \times sens \times prec}{sens + \beta \times prec} \quad (4.6)$$

- *G-mean* (*GSP*): é a média geométrica da sensibilidade e precisão:

$$GSP = \sqrt{sens \times prec} \quad (4.7)$$

- *G-mean2* (*GSS*): é a média geométrica da sensibilidade e especificidade:

$$GSS = \sqrt{sens \times spec} \quad (4.8)$$

- **AUC: Area under the ROC (Receiver operating characteristic) curve** Uma curva ROC é um gráfico bidimensional em que o eixo horizontal representa os valores da taxa de falsos positivos ( $1 - \text{especificidade}$ ) e o eixo vertical a sensibilidade. Um classificador discreto é representado por um ponto no espaço ROC. A sua curva ROC é formada ligando-se o ponto do classificador às extremidades da diagonal, ou seja, aos pontos (0,0) e (1,1), como pode ser visto na Figura 4.1. Uma estratégia para avaliar o desempenho de um classificador é calcular a área sob a curva ROC (AUC) (Figura 4.2). Assim, o classificador que tiver maior área é considerado o melhor.

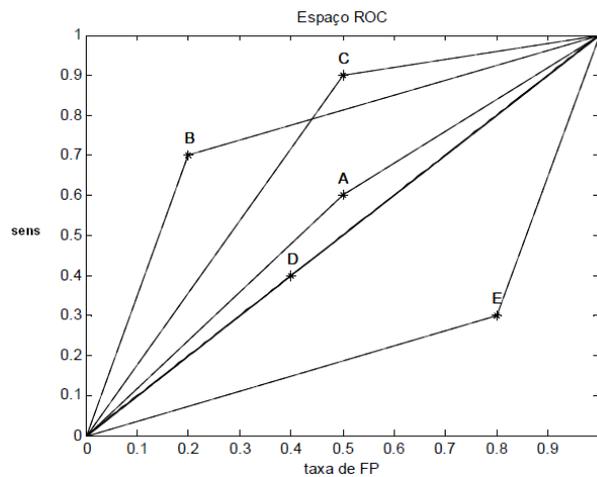


Figura 4.1: Exemplos de curvas ROC para classificadores A, B, C, D e E.

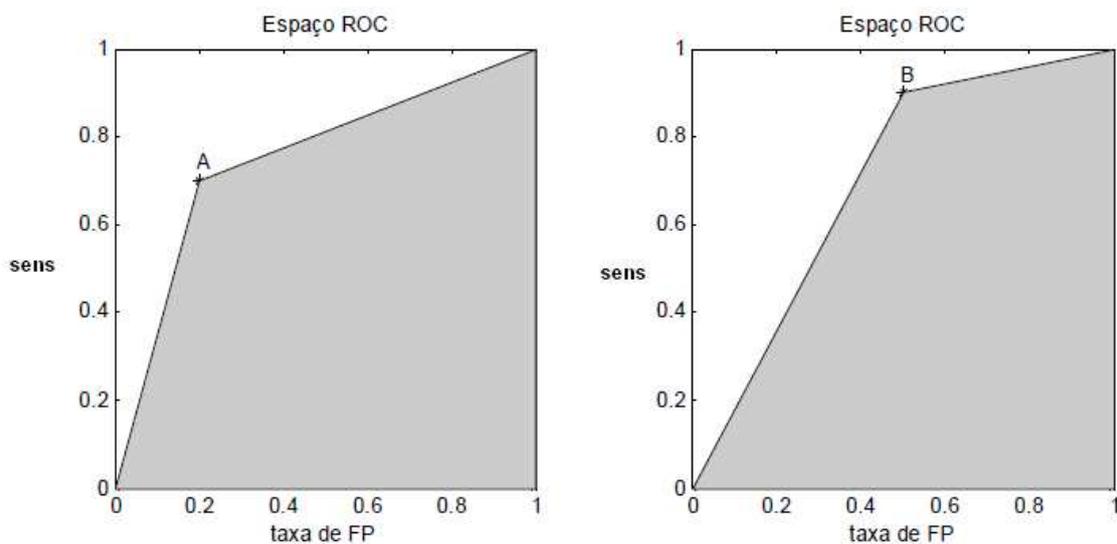


Figura 4.2: Exemplos de AUCs para dois classificadores.

Estendendo-se as métricas anteriores para problemas multi-classes, tem-se:

- TP seriam os positivos verdadeiros de  $c_i (i \in [1, \dots, K])$ ;

- TN quantidade de elementos negativos da classe  $c_k$  ( $k \in [1, \dots, K]$ ) e ( $k \neq i$ ) classificados como negativos;
- FP quantidade de elementos negativos da classe  $c_k$  ( $k \in [1, \dots, K]$ ) e ( $k \neq i$ ) classificados como positivos;
- FN quantidade de elementos da classe  $c_i$ , classificados como  $c_k$  ( $k \in [1, \dots, K]$ ) e ( $k \neq i$ )

$$sens = \sum_{i=1}^K sens(c_i) \times f_r(c_i) \quad (4.9)$$

$$spec = \sum_{i=1}^K spec(c_i) \times f_r(c_i) \quad (4.10)$$

$$prec = \sum_{i=1}^K prec(c_i) \times f_r(c_i) \quad (4.11)$$

$$FMea = \sum_{i=1}^K FMea(c_i) \times f_r(c_i) \quad (4.12)$$

$$GSP = \sum_{i=1}^K GSP(c_i) \times f_r(c_i) \quad (4.13)$$

$$GSS = \sum_{i=1}^K GSS(c_i) \times f_r(c_i) \quad (4.14)$$

$$AUC = \sum_{i=1}^K AUC(c_i) \times f_r(c_i) \quad (4.15)$$

onde:

- $K$  é o número de classes,
- $c_i$  representa  $i$ -ésima classe,
- $f_r(c_i)$  é a frequência relativa da classe  $c_i$ .

## 4.2.2 A Técnica *Bagging*

*Bagging* – acrônimo de **bootstrap aggregating** – é um método que cria múltiplos modelos baseados em dados de treino diferentes e gera um modelo final agregado como sendo um comitê de classificadores. Sua função é reduzir a instabilidade do preditor com relação aos dados selecionados para treino [6]. Em suma, o que este método faz é “empacotar” diversos conjuntos de treinamento distintos entre si, estabelecer um modelo para cada

pacote e calcular a acurácia do preditor como sendo o resultado da votação majoritária dos modelos.

Há, também, como determinar outros dados estatísticos importantes, como o melhor e o pior pacote para treino e o grau de sensibilidade do classificador com relação à escolha dos exemplos de treinamento.

A seguir, serão vistas em detalhes as etapas constituintes do *Bagging*.

### ***Bootstrapping***

O *bootstrapping* é o método de seleção de amostra de dados para cada iteração do *bagging*. Ou seja, cada classificador membro do comitê é treinado por um conjunto de dados diferente, gerado por esta técnica. A distribuição por *bootstrap* é realizada de maneira simples. Suponha que se deseja criar  $n$  pacotes de treinamento contendo  $k$  amostras. O pseudo-código para geração destes pacotes pode ser visto no Algoritmo 3.

---

#### **Algoritmo 3** Algoritmo *Bootstrapping*

---

◊ **Entrada:** Porcentagem da base de dados separada para treino *DadosTreino*,  $n$  número de pacotes e  $k$  número de amostras  
◊ **Saída:**  $n$  pacotes para geração do comitê dos classificadores  
**para cada**  $i = 1$  até  $n$  **faça**  
    **para cada**  $j = 1$  até  $k$  **faça**  
        ◊  $PacoteTreino_i = DadosTreino[aleatório(0, k - 1)]$  /\* Repare que pode haver não seleção ou repetição de amostras \*/  
    **fim para**  
**fim para**

---

Nesta técnica, a construção do conjunto para treinar um classificador permite que sejam selecionados dados repetidos e que certos registros não sejam escolhidos. Isto permite que cada pacote explore, na construção do classificador, uma determinada porção do espaço de busca.

### **Formação dos Membros do Comitê Classificador**

Para a formação dos  $n$  classificadores membros do comitê, executa-se o algoritmo de aprendizado para cada um dos  $n$  pacotes obtidos por meio do *bootstrapping*.

Obtidos os classificadores, a base de teste é aplicada a cada membro classificador e a classe predita é determinada pela votação majoritária do comitê. Por exemplo, suponha que se deseja saber se a amostra de dados cujo valor 13 é maior ou não que 10. Suponha que você tenha 3 votantes (classificadores), A, B, C. A diz que é Maior, B diz que é Menor e C diz que é Maior. A predição final será que a amostra de valor 13 é da classe “Maior” e, portanto, acertou esta amostra de teste.

### 4.2.3 SMOTE *Synthetic Minority Over-sampling Technique*

É uma técnica de amostragem na qual a classe minoritária é aumentada pela criação de exemplos sintéticos [9]. Os novos exemplos são gerados através de perturbações nos dados de treinamento, por meio de operações realizadas no espaço de atributos dos exemplos de treino. A classe com menos exemplos é aumentada tomando cada um de seus exemplos e introduzindo dados sintéticos ao longo do segmento de linha que junta todos os  $k$  vizinhos mais próximos da classe. Dependendo da quantidade de aumento necessária, um subconjunto dos  $k$  vizinhos é escolhido aleatoriamente. Por exemplo, em um aumento de 200%, apenas dois dos  $k$  vizinhos serão escolhidos, e uma amostra é gerada na direção de cada um.

As amostras sintéticas são geradas da seguinte maneira: toma-se a diferença entre o vetor de atributos da amostra em consideração e o seu vizinho mais próximo. Multiplica-se essa diferença por um número aleatório no intervalo  $[0,1]$  e adiciona-se este valor ao vetor original. Isto causa a seleção de um ponto aleatório ao longo do segmento de linha entre dois atributos específicos. O pseudo-código do algoritmo pode ser visto no Algoritmo 4.

Os resultados encontrados em [9] são consequência da combinação do SMOTE com a redução dos exemplos das classes com número maior de amostras. Esta diminuição do tamanho da classe é feita selecionando-se aleatoriamente um exemplo e eliminando o mesmo da base. Esta segunda técnica não foi adotada nos experimentos, pois acredita-se que o mecanismo supressor proposto seja capaz de produzir reduções mais direcionadas.

### 4.2.4 O Problema de Exemplos Irrelevantes

Assim como certos atributos têm mais importância que outros no processo de classificação [5, 40], há também exemplos com essas mesmas características. Algumas pesquisas na busca das instâncias que melhor representem uma classe assumem a presença de um “tutor” que apresenta ao algoritmo de aprendizado instâncias informativas e seqüências ideais de treinamento. Entretanto, uma abordagem mais eficaz envolve a seleção dos casos de treino pelo classificador.

Há três razões principais para a seleção de exemplos de treinamento [5]. Caso haja um número de registros suficiente, faz sentido aprender a partir daqueles mais relevantes, para aumento de desempenho computacional. O segundo motivo é quando há um custo associado à definição dos rótulos dos exemplos e há muitas instâncias disponíveis. Finalmente, a terceira razão é para aumentar a taxa de treinamento focando a atenção nos exemplos mais informativos, o que ajuda na busca através do espaço de hipóteses.

Há também que se fazer uma distinção entre relevância de exemplos do ponto de vista da informação que ele carrega ou aqueles que são relevantes para suprir a demanda de treinamento de um determinado algoritmo. A maior parte das pesquisas atuais se foca na

---

**Algoritmo 4** Algoritmo SMOTE( $T, N, k$ )

---

◇ **Entrada:** Número de exemplos  $T$  da classe minoritária; Porcentagem  $N$  de aumento da classe; Número de vizinhos mais próximos  $k$

◇ **Saída:**  $(N/100) * T$  exemplos sintéticos

◇ /\*Se  $N$  for menor que 100, apenas uma porcentagem da classe será trabalhada\*/

**se**  $N < 100$  **então**

    ◇  $T = N/100 * T$  /\*Parte da classe\*/

    ◇  $N = 100$

**fim se**

◇  $N = (int)N/100$

**para cada**  $i = 1$  até  $T$  **faça**

    ◇ Computar os  $k$  vizinhos mais próximos de  $i$  e armazenar em  $vetKNN$

**enquanto**  $N \neq 0$  **faça**

        ◇  $nn = aleatório(1, k)$

        ◇  $n =$  Número de Atributos

**para cada**  $j = 1$  até  $n$  **faça**

            ◇  $diferenca = Amostras[vetKNN[nn]][j] - Amostras[i][j]$

            ◇  $espacamento = aleatorio(0, 1)$

            ◇  $Sinteticos[novo][j] = Amostras[i][j] + espacamento \times diferenca$

**fim para**

        ◇  $novo++$

        ◇  $N--$

**fim enquanto**

**fim para**

---

seleção de dados voltada para o classificador.

Em [5], os esquemas de seleção de exemplos de treinamento são divididos em dois conjuntos, o primeiro escolhe instâncias a partir de exemplos rotulados e o segundo a partir de casos não rotulados. O primeiro caso, que representa o problema abordado neste trabalho – pois trata-se de classificação, será visto a seguir.

### **Seleção de Exemplos Rotulados**

A primeira abordagem assume que um conjunto de dados rotulados está disponível para uso no aprendizado, porém nem todos os exemplos são úteis. Conforme citado anteriormente, pode-se embutir a seleção dos exemplos de treinamento no algoritmo de classificação. Por exemplo, Littlestone e Mesterharm [49] mostraram que uma variante do *naive Bayes* que aprende apenas a partir de erros lida melhor com exemplos sem significância do que a versão clássica, que atualiza sua estatística a cada exemplo.

Se teste e treino são ambos gerados sob a mesma distribuição, então há uma grande probabilidade de que os dados de treinamento serão como um todo relevantes para o aprendizado. Contudo, com o progresso do aprendizado, o conhecimento do classificador passa a ser mais relevante que alguns dados.

Em certos modelos, os algoritmos de treinamento precisam dobrar o número de exemplos a serem apresentados de forma que diminuam a taxa de erro pela metade. Entretanto, para métodos mais tradicionais, o número de exemplos é sempre proporcional à taxa de erro. Para lidar com este tipo de aprendizado, Schapire [64] descreve um método que, a partir de um aprendizado genérico, remove alguns dados de treinamento baseados no comportamento do algoritmo.

Um outro método chamado *windowing*, concebido por Quinlan [59], tenta reduzir o tempo de construção de árvores de decisão a partir de bases de dados muito grandes. Seleciona-se uma mostra de dados aleatória para se construir uma árvore inicial que irá classificar os dados restantes. A partir dos exemplos classificados incorretamente, o método seleciona outro conjunto aleatório para se unir ao primeiro conjunto. Quinlan demonstrou que há ganho no desempenho da construção da árvore de decisão.

Os próximos capítulos apresentam as técnicas desenvolvidas neste trabalho para lidar com o problema de redução do conjunto de treinamento. O capítulo a seguir trata de um algoritmo de classificação cuja inspiração provém do funcionamento de certas partes dos sistemas imunitários biológicos. Em seguida, apresenta-se um mecanismo de supressão desenvolvido para atuar juntamente com alguns tipos de classificadores.

# Capítulo 5

## O Algoritmo SUPRAIC (*Suppressor Artificial Immune Classifier*)

### 5.1 Introdução

Classificação [31] e reconhecimento de padrões são tarefas importantes em todos os campos de conhecimento [22]. Classificar significa categorizar, em um finito número de classes, elementos definidos por um grupo de atributos.

Há muitos tipos de classificadores estatísticos e baseados em inteligência artificial, como pode ser visto em [34, 52]. Dentre as principais questões em problemas de classificação está a escolha de amostras significativas para o treinamento supervisionado. Um conjunto de treinamento capaz de representar bem as características de uma classe provê boas chances de se estabelecer um preditor eficaz.

Neste capítulo é proposto um classificador inspirado em certos comportamentos dos sistemas imunológicos [25]. Sua característica principal comparadas a outros classificadores imunes artificiais, que podem ser vistos em [80, 86, 85, 82, 30, 84, 83, 81, 51, 8, 73, 72, 51, 19, 74, 46, 76, 47, 75], é a premissa de que o próprio conjunto de treinamento constitui a população inicial de anticorpos do sistema. Há também um mecanismo de supressão que tenta reduzir o conjunto inicial de treinamento em um subconjunto contendo menos registros. Acredita-se que tal subconjunto contenha os exemplos mais relevantes, sem que haja perdas significativas na capacidade de generalização do preditor.

Outra característica deste mecanismo que o difere dos demais citados é a sua simplicidade. Não existem artifícios como maturação de afinidade, seleção clonal [16, 17], e não é baseado em modelos de redes imunológicas ou redes de *Kohonen* [46]. As vantagens e problemas desta nova concepção serão discutidos à frente.

As seções deste capítulo estão organizadas da seguinte forma. A próxima explica em detalhes os fenômenos biológicos que inspiraram o algoritmo. Na Seção 5.2 apresenta-se uma descrição detalhada do algoritmo proposto juntamente com as primeiras idéias rela-

cionadas ao processo de supressão. Experimentos e resultados obtidos a partir de bases de dados encontradas na literatura são expostos na Seção 5.3. Na Seção 5.4, encontram-se as considerações finais sobre o modelo e as motivações para seu melhoramento que produziram a nova abordagem do capítulo seguinte.

### 5.1.1 O Comportamento dos Sistemas Imunitários Biológicos

De acordo com o mecanismo de auto-regulação dos SIs, clones que não são mais necessários para o organismo, ou aqueles que são auto-ímmunes, não recebem estímulos de sobrevivência e sofrem apoptose. Estes sinais são emitidos pelos linfonodos<sup>1</sup> e células T Auxiliares<sup>2</sup> [38]. Esta característica de auto-regulação impede o desperdício de energia dentro do organismo e faz com que seja mantido apenas o repertório de linfócitos realmente necessário para defesa contra patógenos. Estes conceitos apresentados serviram como inspiração para o algoritmo apresentado na próxima seção (Seção 5.2).

Basicamente, o algoritmo apresentado parte do conceito no qual o modelo para o sistema deve evoluir de forma a produzir anticorpos que reconheçam o conjunto de treinamento e sejam capazes de identificar novos antígenos apresentados. Ao invés de trabalhar como um sistema que gera e evolui clones de células B até que os anticorpos reconheçam o conjunto de treinamento, como feito em [80, 86, 85, 82, 30, 84, 83], esta nova abordagem propõe que os dados de treinamento por si só sejam o repertório de anticorpos do sistema.

O conceito de supressão é empregado no conjunto de treinamento para eliminar anticorpos muito parecidos. Para que haja supressão, os antígenos, ou dados de teste, são divididos em dois subgrupos. O primeiro será responsável por testar o modelo e eliminar anticorpos redundantes. O segundo subgrupo é usado para validar a eficácia dos anticorpos restantes após a supressão.

## 5.2 O Algoritmo Proposto

O algoritmo parte da idéia de que o modelo do sistema deve produzir anticorpos que reconheçam o conjunto dos dados de treinamento e esteja apto a identificar novos antígenos apresentados. Portanto, ao invés do sistema gerar e evoluir clones de células B até que os anticorpos reconheçam o conjunto de treinamento e definam-se as células de memória, é proposto que o conjunto de treino constitua o repertório de anticorpos do sistema. Isto quer dizer que a memória imunológica, representada por *Artificial Recognition Balls* (ARBs)<sup>3</sup>, que são metáforas matemáticas para anticorpos ou células B [57, 58, 17], (ver

---

<sup>1</sup>Nos linfonodos se concentram as fontes desses estímulos, e concomitantemente, as respostas a esses estímulos são coordenadas pelos linfócitos T.

<sup>2</sup>*T helper cells*

<sup>3</sup>Bolas de Reconhecimento Artificial.

Figura 5.1), estará pronta para generalização a partir do momento que os dados de treinamento são inseridos no sistema. Na figura apresentada,  $S$  é o chamado *shape space* onde  $\bullet$  são os anticorpos,  $\times$  são os antígenos e  $\varepsilon$  denota o raio de cada anticorpo [58, 17].

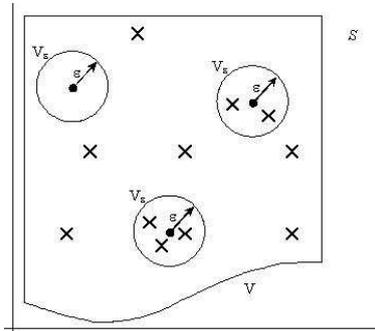


Figura 5.1: O modelo *Shape Space*.

O conjunto de atributos é normalizado para pertencer à mesma escala de valores. Neste caso, os dados foram mapeados para o intervalo  $[0,1]$ . Foram considerados atributos dos tipos inteiro e real.

Para o aprendizado, a base de dados é dividida em três subconjuntos, um para treino, um para supressão e outro de teste ou validação. A proporção inicial das amostras adotada para cada grupo foi de, respectivamente, 60%, 20% e 20%. Em um outro experimento, foram utilizadas as proporções de 70%, 20% e 10%. Estes valores foram adotados de forma empírica, baseados em números que são usualmente empregados.

Os antígenos e anticorpos são representados por um vetor de atributos. Os antígenos, ou dados de validação, são classificados de acordo com o anticorpo mais próximo. Isto significa que, neste caso, ao invés de um *shape space* com  $\varepsilon$  variante, o modelo pode ser representado pelo diagrama de Voronoi [79] para  $n$  dimensões, como pode ser visto na Figura 5.2. Na figura, os pontos são os centróides das ARBs e representam os anticorpos que cobrem um certo grupo de antígenos mais próximos.

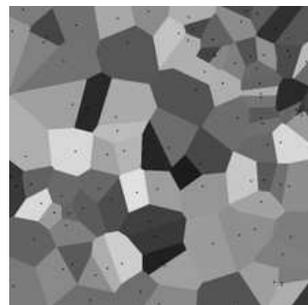


Figura 5.2: Diagrama de Voronoi: uma representação do processo de classificação.

O anticorpo mais próximo é determinado por uma medida de distância, assim como todos os algoritmos da classe dos  $K$  vizinhos mais próximos (*K Nearest Neighbours* –

KNN). Nos experimentos foram adotadas as distâncias Euclidiana e de *Manhatan*, descritas, respectivamente pelas equações 5.1 e 5.2.

$$distanciaEuclidiana = \sqrt{\sum_{i=1}^{TotalAtributos} (Anticorpo[i] - Antigeno[i])^2} \quad (5.1)$$

$$distanciaManhattan = \sum_{i=1}^{TotalAtributos} |Anticorpo[i] - Antigeno[i]| \quad (5.2)$$

A razão pela qual os anticorpos (conjuntos de supressão e treino) foram separados em dois subconjuntos é dada pelo mecanismo de supressor. Ao conjunto de treinamento é aplicada a supressão de forma que anticorpos muito similares sejam eliminados e só os melhores sejam mantidos. Em outras palavras, os anticorpos capazes de reconhecer antígenos a partir do conjunto de supressão permanecem no organismo, enquanto os demais são eliminados da população.

Neste classificador artificial, os sinais de sobrevivência para cada anticorpo são representados por uma variável contadora que armazena o número de antígenos reconhecidos, independente se a classificação pelo anticorpo mais próximo está correta ou não. Sabe-se que algumas bases de dados representam problemas de classificação mais difíceis, com uma área de similaridade entre duas ou mais classes grandes, o que faz com que a maioria dos classificadores apresente falhas nestas regiões. Por esta razão, decidiu-se manter os anticorpos que contêm instâncias classificadas erroneamente na população já suprimida. Estes anticorpos podem ser vistos sob a perspectiva biológica como aqueles que apresentam reação cruzada <sup>4</sup>.

O esquema do modelo pode ser visto no Algoritmo 5, chamado SUPRAIC <sup>5</sup> [25].

## 5.3 Experimentos e Resultados

Nesta seção são apresentados experimentos utilizando bases de dados padrão extraídas do repositório de aprendizado de máquina da Universidade da Califórnia, Irvine [20]. As métricas adotadas para avaliar a eficiência do classificador foram aquelas apresentadas no capítulo anterior e em [22].

### 5.3.1 Base de Dados Câncer

O primeiro experimento foi realizado sobre a base de dados *Breast Cancer*. A Tabela 5.1 mostra os resultados do classificador sem utilizar o mecanismo de supressão proposto. Já na Tabela 5.2, o mecanismo de supressão foi aplicado.

<sup>4</sup>Cross reactivity

<sup>5</sup>Suppressor Artificial Immune Classifier.

---

**Algoritmo 5** Algoritmo SUPRAIC

---

- ◊ Ler o arquivo contendo a base de dados;
  - ◊ Normalizar os dados dentro do intervalo [0,1];
  - ◊ Determinar anticorpos e antígenos dividindo os dados (distribuídos uniformemente) em, por exemplo, 60% anticorpos e 20% supressão e 20% antígenos;
  - ◊ Para cada anticorpo, inicializar sua variável contador = 0;
  - ◊ Treinar o modelo de anticorpos com o conjunto supressor encontrando o anticorpo mais próximo a cada antígeno. O anticorpo mais próximo é medido pela distância entre os atributos;
  - ◊ Para cada antígeno reconhecido pelo anticorpo, aumentar em um o valor contido na variável contador;
  - ◊ Suprimir os demais anticorpos incapazes de reconhecer antígenos (contador = 0). Eliminá-los da população de anticorpos;
  - ◊ Validar o preditor final, já suprimido pelo passo anterior, utilizando os anticorpos restantes para reconhecer o conjunto de validação;
  - ◊ Calcular a acurácia do preditor final.
- 

Comparando os resultados nas Tabelas 5.1 e 5.2, não há muita diferença entre os valores das métricas comparando os dados de treino no seu estado inicial ou suprimidos. Entretanto, a quantidade de anticorpos foi reduzida de 419 para 56. Isto significa uma diminuição de quase 87% dos dados de treinamento. A Figura 5.3 mostra sob a forma de coordenadas estrela os dados de treinamento antes da supressão e a Figura 5.4 mostra o classificador final já com os dados suprimidos.

porcentagens	distância	acc	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidiana	0.97	0.94	0.99	0.98	0.48	0.96	0,96
60-20-20	Manhattan	0.98	0.94	1.00	1.00	0.47	0.95	0.95
70-20-10	Euclidiana	0.97	0.93	1.00	1.00	0.48	0.96	0.96
70-20-10	Manhattan	0.94	0.85	1.00	1.00	0.46	0.92	0.92

Tabela 5.1: Resultados do classificador sem utilizar o mecanismo de supressão para a base de dados Câncer com 419 anticorpos.

porcentagens	distância	acc	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidiana	0.94	0.85	0.99	0.98	0.45	0,91	0,91
60-20-20	Manhattan	0.96	0.88	1.00	1.00	0.47	0.94	0.94
70-20-10	Euclidiana	0.96	0.89	1.00	1.00	0.47	0.94	0.94
70-20-10	Manhattan	0.96	0.88	1.00	1.00	0.47	0.94	0.94

Tabela 5.2: Resultados do classificador utilizando o mecanismo de supressão para a base de dados Câncer com 56 anticorpos.

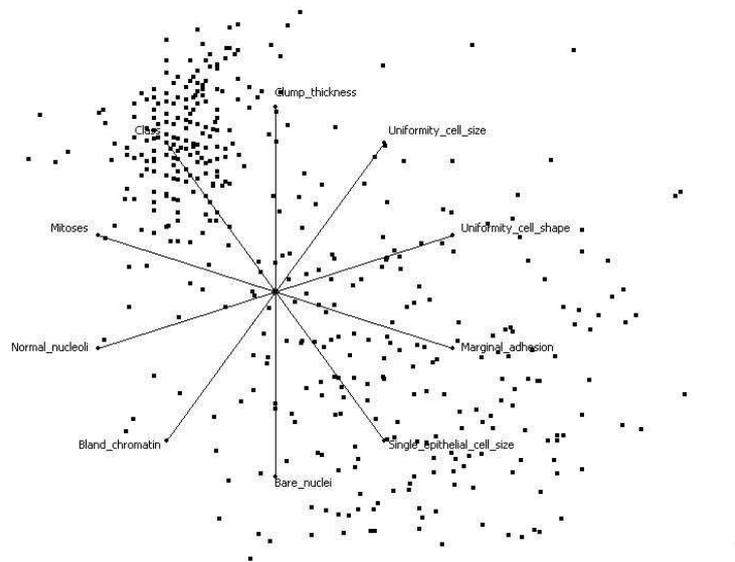


Figura 5.3: Quantidade total de anticorpos (dados de treinamento) para a base de dados Câncer.

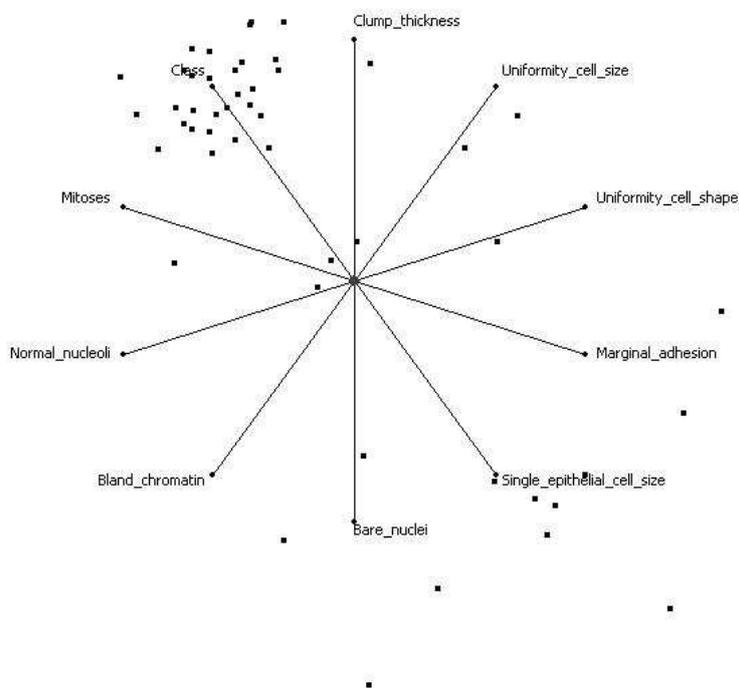


Figura 5.4: Anticorpos da base de dados Câncer após supressão e divisão de porcentagens em 60–40%.

### 5.3.2 Base de Dados Diabetes dos Índios Pima

A base de dados Diabetes contém 768 registros, 8 atributos e 2 classes. Assim como na primeira base de dados estudada, pode-se ver que os resultados das tabelas 5.3 e 5.4 não são muito diferentes, exceto pela métrica especificidade. Mesmo assim, a redução dos anticorpos plea supressão foi de 460 para 80. As Figuras 5.5 e 5.6 mostram, respectivamente, os dados de treinamento antes e depois da supressão.

porcentagens	distância	acc	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidiana	0.67	0.7	0.62	0.77	0.36	0.73	0.66
60-20-20	Manhattan	0.69	0.71	0.65	0.79	0.37	0.75	0.68
70-20-10	Euclidiana	0.69	0.74	0.61	0.74	0.37	0.74	0.67
70-20-10	Manhattan	0.71	0.74	0.68	0.77	0.38	0.76	0.71

Tabela 5.3: Resultados do preditor sem mecanismo de supressão para base de dados Diabetes com 460 anticorpos.

porcentagens	distância	acc	sens	spec	prec	Fmea	GSP	GSS
60-20-20	Euclidiana	0.65	0.77	0.45	0.71	0.37	0.74	0.59
60-20-20	Manhattan	0.65	0.8	0.4	0.7	0.37	0.74	0.56
70-20-10	Euclidiana	0.65	0.72	0.55	0.70	0.35	0.71	0.63
70-20-10	Manhattan	0.66	0.80	0.45	0.68	0.37	0.74	0.60

Tabela 5.4: Resultados do preditor com mecanismo de supressão para base de dados Diabetes com 80 anticorpos.

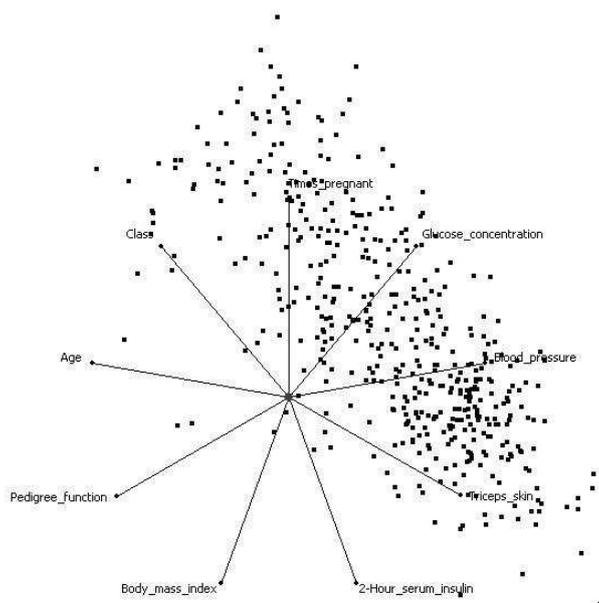


Figura 5.5: A quantidade total de anticorpos (dados de treinamento) para a base de dados Diabetes.

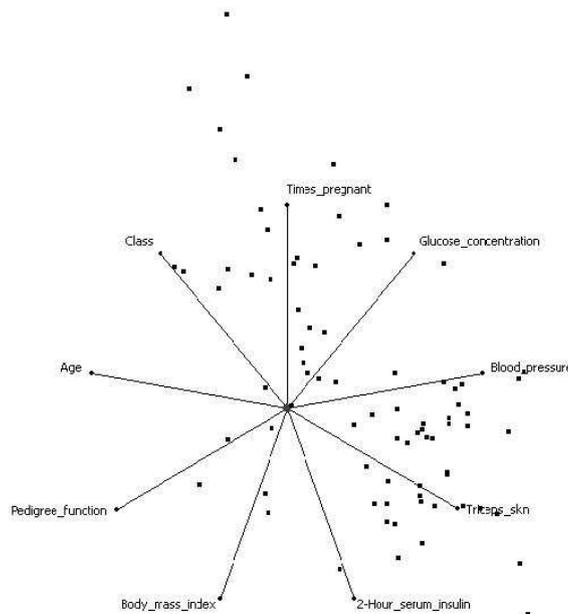


Figura 5.6: Anticorpos da base de dados Diabetes depois da supressão com divisão de porcentagens em 60–40%.

### 5.3.3 Resultados para Outras Bases de Dados Estudadas

Além das bases anteriores, o algoritmo SUPRAIC foi rodado e comparado a outros métodos. Para os estudos realizados, foram implementados em ambiente Matlab/C++ os seguintes classificadores:

1. KNN - Método dos  $k$  vizinhos mais próximos com distância Euclidiana [34];
2. Discriminate linear [63, 34];
3. Árvore de decisão - J4.8 [63, 59, 34];
4. SVM - Máquina de vetor de suporte com *kernel* de função de base radial [11];
5. Classificador *Bayesiano* [63, 34];
6. Classificador *Fuzzy* com regras de Sugeno [61, 41, 45];
7. MLP (*multilayer perceptron*) com 3 camadas utilizando *backpropagation* [35, 63, 59, 34].

As bases testadas foram: *Monks-2* (Tabela 5.5), *Monks-3* (Tabela 5.6), *Ionosphere* (Tabela 5.7), *Well* (Tabela 5.8), *Contraceptivos* (Tabela 5.9) e *Wine* (Tabela 5.10). A caracterização destas bases, juntamente com outras que serão estudadas no próximo capítulo podem ser vistas no Anexo A. Os resultados coletados foram obtidos por meio da média de cem

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.70	0.77	0.55	0.77	0.38	0.77	0.65	0.66	59
10NN	0.78	0.79	0.76	0.91	0.42	0.85	0.77	0.72	360
DL	0.65	0.66	0.07	1.00	0.39	0.81	0.09	0.50	360
AD	0.84	0.88	0.77	0.87	0.44	0.88	0.82	0.82	360
SVM	0.85	0.92	0.75	0.85	0.44	0.88	0.83	0.85	360
<i>Fuzzy</i>	0.67	0.67	0.68	0.97	0.40	0.81	0.66	0.53	360
<i>Bayes</i>	0.74	0.87	0.59	0.71	0.39	0.78	0.72	0.76	360
MLP	0.78	0.83	0.69	0.84	0.42	0.84	0.76	0.76	360

Tabela 5.5: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Monks-2*.

rodadas com conjuntos inicializados, através de sementes diferentes para o gerador de números pseudo-aleatório, para todos os modelos. A distância utilizada foi Euclidiana. Uma outra alternativa, não contemplada neste estudo, seria implementar o *cross validation*.

Os resultados mostrados na Tabela 5.5 mostram que para a base *Monks-2* a classificação pelo algoritmo SUPRAIC em média produz resultados positivos, sendo que este algoritmo superou o desempenho dos classificadores DL e *Fuzzy* para a métrica de acurácia.

Já para a base *Monks-3* o resultado, comparado aos demais algoritmos de classificação foi inferior, como pode ser visto na Tabela 5.6. Uma explicação para este fato se deve à baixa tolerância a ruídos característica do algoritmo proposto. Se houver exemplos ruidosos dentro do conjunto de treinamento ou supressão, a determinação do classificador final pode ser comprometida.

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.74	0.71	0.77	0.74	0.36	0.72	0.73	0.74	61
10NN	0.89	0.85	0.94	0.94	0.44	0.89	0.89	0.89	332
DL	0.77	0.77	0.78	0.76	0.38	0.76	0.77	0.77	332
AD	0.99	0.99	0.98	0.98	0.49	0.99	0.99	0.99	332
SVM	0.96	0.96	0.97	0.96	0.48	0.96	0.96	0.96	332
<i>Fuzzy</i>	0.68	0.67	0.71	0.69	0.34	0.67	0.68	0.69	332
<i>Bayes</i>	0.91	0.85	0.99	0.99	0.46	0.91	0.91	0.91	332
MLP	0.98	0.97	0.98	0.98	0.49	0.97	0.98	0.98	332

Tabela 5.6: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Monks-3*.

Nas bases de dados *Ionosphere* (Tabela 5.7), *Well* (Tabela 5.8), *Contraceptivos* (Tabela 5.9), o algoritmo é eficiente, sendo algumas vezes superior ou equivalente aos demais classificadores.

Para a base de dados *Wine* (Tabela 5.10) o resultado foi bastante inferior comparado aos demais métodos. Apesar dos resultados mostrados não serem superiores, mostram

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.81	0.54	0.97	0.91	0.33	0.69	0.71	0.75	40
10NN	0.81	0.78	0.95	0.98	0.43	0.88	0.86	0.74	210
DL	0.86	0.83	0.95	0.98	0.45	0.90	0.89	0.82	210
AD	0.88	0.90	0.84	0.91	0.45	0.90	0.87	0.86	210
SVM	0.94	0.94	0.94	0.97	0.48	0.95	0.94	0.93	210
Fuzzy	0.81	0.78	0.96	0.99	0.43	0.88	0.86	0.74	210
Bayes	0.64	0.64	0.00	1.00	0.39	0.80	0.00	0.50	210
MLP	0.88	0.87	0.93	0.97	0.46	0.91	0.89	0.85	210

Tabela 5.7: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Ionosphere*.

uma potencial utilização desta idéia (que será mostrada no próximo capítulo). O desempenho do classificador SUPRAIC foi inferior, porém trabalhando com um conjunto de dados para treinamento bastante menor.

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.67	0.67	0.78	0.68	0.33	0.67	0.72	0.72	73
10NN	0.76	0.76	0.85	0.77	0.38	0.76	0.80	0.80	903
DL	0.76	0.76	0.85	0.77	0.38	0.76	0.80	0.80	903
AD	0.77	0.77	0.86	0.78	0.39	0.77	0.81	0.82	903
SVM	0.76	0.76	0.85	0.77	0.38	0.76	0.80	0.81	903
Fuzzy	0.70	0.70	0.81	0.71	0.35	0.70	0.75	0.75	903
Bayes	0.76	0.76	0.85	0.77	0.38	0.76	0.80	0.81	903
MLP	0.49	0.49	0.40	0.91	0.31	0.65	0.34	0.44	903

Tabela 5.8: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Well*.

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.40	0.54	0.24	0.46	0.25	0.50	0.35	0.39	76
10NN	0.49	0.49	0.63	0.50	0.25	0.49	0.55	0.56	883
AD	0.50	0.50	0.64	0.51	0.25	0.50	0.56	0.57	883
DL	0.50	0.50	0.64	0.54	0.26	0.52	0.56	0.57	883
SVM	0.49	0.49	0.63	0.60	0.26	0.53	0.55	0.56	883
Fuzzy	0.45	0.45	0.59	0.48	0.23	0.46	0.51	0.52	883
Bayes	0.49	0.49	0.70	0.56	0.24	0.51	0.57	0.60	883
MLP	0.53	0.53	0.67	0.56	0.27	0.54	0.59	0.60	883

Tabela 5.9: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Contraceptivos*.

Classificador	acc	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
SUPRAIC	0.65	0.65	0.51	0.51	0.28	0.57	0.55	0.58	12
10NN	0.93	0.93	0.96	0.94	0.47	0.94	0.95	0.95	106
DL	0.99	0.99	0.99	0.99	0.49	0.99	0.99	0.99	106
AD	0.91	0.91	0.95	0.92	0.45	0.91	0.93	0.93	106
SVM	0.97	0.97	0.98	0.97	0.48	0.97	0.97	0.97	106
<i>Fuzzy</i>	0.93	0.93	0.96	0.93	0.46	0.93	0.94	0.94	106
<i>Bayes</i>	0.98	0.98	0.98	0.98	0.49	0.98	0.98	0.98	106
MLP	0.98	0.98	0.99	0.98	0.49	0.98	0.98	0.98	106

Tabela 5.10: Comparativo dos resultados do classificador SUPRAIC com outros algoritmos de classificação para a base de dados *Wine*.

## 5.4 Considerações a Respeito do Modelo

Este capítulo apresentou a proposta de um método classificador de dados determinístico da classe dos KNN, utilizando um mecanismo supressor inspirado na conduta dos SIS biológicos. Particularmente, a idéia surgiu da observação do comportamento de células T auxiliares e linfonodos mediante clones de linfócitos que não são mais necessários para o organismo, ou que são auto-reativos.

Basicamente, o preditor trabalha com uma base de dados dividida em três partes. A primeira representa os anticorpos e contém a maior parte dos dados, de forma que, teoricamente, possa cobrir todo o espaço de classificação. A segunda é o chamado conjunto supressor, responsável por ajudar o sistema a eliminar os anticorpos que não foram utilizados. A supressão é realizada partindo do princípio de que os anticorpos capazes de reconhecer algum antígeno devem permanecer na população. Os demais são suprimidos – retirados da população. A terceira parte representa o conjunto de validação. Seu papel é certificar a eficácia do trabalho realizado pelo mecanismo supressor.

Experimentos realizados utilizando alguns *benchmarks* contendo duas ou mais classes apresentaram resultados satisfatórios para algumas bases. Como os resultados foram obtidos a partir de fontes públicas de dados, sem focar um problema específico, a representação do anticorpo e o *bias* indutivo foi o mesmo para todos os casos estudados, contrariando as idéias de se propor soluções específicas para problemas de classificação [27, 87].

O sistema apresenta alguns pontos negativos. Apesar de seu objetivo principal ser se comportar como uma meta-heurística que trabalha no sentido de privilegiar os dados mais representativos dentro do conjunto de treinamento, seu caráter determinístico não permite maleabilidade no seu processo de aprendizado. A solução será sempre a mesma, e não há a preocupação em evoluir possíveis anticorpos que poderiam ser eficientes e não foram contemplados pelas instâncias do conjunto de treinamento.

Sob o ponto de vista da descoberta do conhecimento, e em particular a classificação, o sistema tende a não detectar anomalias [3, 4, 14, 21, 23, 29, 32, 33, 44, 48, 50, 56, 65,

66, 68, 69, 70, 88, 89, 90]. Exemplos observados seriam casos raros que ocorrem dentro de certas classes. Este tipo de dado é tratado pelo modelo, quando anticorpo, como um imunógeno ineficaz, ou quando antígeno, como um dado classificado incorretamente.

Logo, neste caso, não se contempla a predição onde dados fora do padrão aparecem. E o objetivo, neste trabalho, é propor técnicas capazes de tratar casos raros e classes raras dentro de problemas de categorização. Para que isto seja possível, faz-se necessária a adoção de algumas técnicas de aprendizado que permitam que o sistema ativamente insira novas instâncias, criadas sinteticamente, dentro da população de anticorpos.

No capítulo seguinte, apresenta-se um aprimoramento desta primeira idéia, já não mais objetivando-se classificar dados, mas obter uma base de treinamento suprimida que contenha os dados mais significativos dentro de todo o conjunto de treino. Este conjunto obtido é testado em alguns classificadores mais comuns, em que cujas características se enquadram esta nova abordagem de treinamento. Igualmente, tenta-se lidar com problemas envolvendo bases desbalanceadas e multi-classes.

# Capítulo 6

## Mecanismo Seletor de Dados Baseado em Imuno-supressão

### 6.1 Introdução

Como foi dito no Capítulo 4, existe um grande interesse na obtenção de métodos eficientes para seleção de exemplos de treinamento [5]. Esta seleção é muito útil em casos onde há bases de dados com um número de registros grande e com exemplos redundantes – quando se deseja aumentar o desempenho computacional. Outro motivo é quando há um custo associado à definição dos rótulos dos exemplos e há muitas instâncias disponíveis. Finalmente, a terceira razão é para se focar nos exemplos mais informativos, ajudando na busca através do espaço de hipóteses.

Neste capítulo será apresentado um mecanismo supressor para ser hibridizado a métodos de classificação de modo que se aumente o desempenho na construção do modelo preditor, sem que isto acarrete em perdas na sua acurácia. Propõe-se, assim, um método de seleção de dados de treinamento. Esta abordagem foi inspirada no comportamento do sistema imunológico.

O capítulo foi organizado da seguinte maneira. A próxima seção (6.2) trata da motivação e funcionamento básico do mecanismo, que surgiu como uma evolução do classificador apresentado no Capítulo 5. Em seguida, na Seção 6.3 é descrito o funcionamento do método de seleção de dados de treinamento e as técnicas utilizadas para lidar com bases desbalanceadas, ruidosas e multi-classes. Na Seção 6.4 são discutidos os resultados conseguidos na classificação das bases de dados utilizando a técnica de *bagging* juntamente com a seleção de treino. Em 6.5 são analisados os resultados com o algoritmo de geração de amostras sintéticas SMOTE. Finalmente, na Seção 6.6, são apresentadas as conclusões a respeito do método.

## 6.2 O Mecanismo Supressor para a Seleção de Dados de Treinamento (*SeleSup*)

No capítulo anterior, foi apresentada uma proposta de classificador inspirada na observação do comportamento de células T auxiliares e linfonodos mediante clones de linfócitos não mais necessários ao organismo.

O classificador trabalha com uma base de dados dividida em três partes, anticorpos, supressão e antígenos. A supressão é realizada partindo do princípio de que os anticorpos capazes de reconhecer algum antígeno devem permanecer na população. Os demais são eliminados. Os resultados apresentados, no geral, foram satisfatórios. Entretanto, foram detectados alguns problemas que serviram como base para a proposta e discussões deste capítulo.

O sistema classificador apresenta alguns pontos negativos, como a falta de maleabilidade no processo de supressão, dando ao algoritmo um caráter determinístico com relação ao conjunto de anticorpos final após a supressão. Isto quer dizer que a solução será sempre a mesma, mantendo-se os conjuntos. Não há, também, a preocupação em gerar possíveis anticorpos que poderiam ser eficientes e não foram contemplados pelas instâncias do conjunto de treinamento. Por exemplo, casos raros que ocorrem dentro de certas classes são ignorados pela abordagem.

Outro problema com o classificador é que, pelo fato de seu aprendizado se dar através da observação de exemplos individualmente, a inserção de ruídos na base pode comprometer a produção do conjunto final de anticorpos. Estes problemas tornam o classificador ineficaz em relação a bases ruidosas, desbalanceadas e com multi-classes.

Apesar dos problemas citados, a idéia da construção de um classificador que tivesse em si embutida uma técnica de seleção dos dados para treino pareceu promissora. Desta forma, o foco da pesquisa foi desviado da criação de um classificador para o desenvolvimento de uma estratégia que pudesse analisar as características do conjunto de treinamento e escolher, dentre as amostras, aquelas que fossem mais representativas e apresentá-las ao método classificador. Preocupou-se, também, em se gerar um algoritmo genérico, que pudesse funcionar produzindo bons resultados para diversos tipos de classificadores. Esquemáticamente, o método pode ser representado pela figura 6.1.

Para se conceber tal método, consideraram-se algumas restrições. Primeiramente, a razão pela qual se deseja selecionar dados de treinamento é manter o desempenho e diminuir o custo computacional. Portanto, neste contexto, não faz sentido a criação de um algoritmo que demande diversas iterações, como por exemplo os algoritmos evolucionários ou algoritmos combinatórios. Uma solução seria desenvolver um método guloso não adaptativo que apresentasse bons resultados.

Para tanto, foi desenvolvido um método de seleção de dados de treinamento com base em interações imunológicas que garantem a sobrevivência dos anticorpos mais adaptados.



Figura 6.1: Esquema do funcionamento do mecanismo de seleção.

O raciocínio para a escolha dos anticorpos, ou dados de treino é semelhante ao definido no capítulo anterior e será descrito na seção a seguir.

### 6.3 O Algoritmo

A seleção dos dados é feita por um mecanismo inspirado no comportamento dos linfócitos e células T com relação a anticorpos pouco adaptados, que são responsáveis por gerar estímulos de sobrevivência para os plasmócitos. De uma forma sucinta, o que acontece no organismo é a manutenção de bons mecanismos de defesa em detrimento de moléculas ineficazes no combate a patógenos.

No contexto da descoberta do conhecimento, para a construção do mecanismo seletor, conjecturou-se que, encarando-se os dados de treinamento como anticorpos e os de teste como antígenos, poderia-se fazer algo semelhante ao mecanismo de supressão imunológico. Isto para se tentar definir, dentre os anticorpos dados, aqueles com maior capacidade de representar o conjunto como um todo.

Para a supressão, a base de dados mais uma vez foi dividida em três subconjuntos, treino, teste e supressão numa proporção de, respectivamente, 60%, 20% e 20%. O conjunto supressor será o responsável por fazer um teste preliminar nos anticorpos, decidindo a aptidão de cada um.

Novamente, os dados são normalizados para o intervalo  $[0,1]$  e antígenos e anticorpos são representados como vetores de atributos em um espaço de  $n$  dimensões.

Os anticorpos devem reconhecer as amostras do conjunto de supressão. O reconhecimento é feito de acordo com o anticorpo mais próximo, via distância Euclidiana. Para

cada antígeno da supressão reconhecido, produz-se um estímulo de sobrevivência para o anticorpo. Sem nenhum estímulo, o anticorpo sai da população. O pseudo-código da supressão é bastante semelhante ao do algoritmo SUPRAIC do capítulo anterior e pode ser visto a seguir, no Algoritmo 6. Este algoritmo é igual ao apresentado no capítulo anterior, porém aplicado somente à seleção de dados.

---

**Algoritmo 6** Algoritmo SeleSup

---

- ◊ Ler o arquivo contendo a base de dados;
  - ◊ Normalizar os dados dentro do intervalo [0,1];
  - ◊ Determinar anticorpos e antígenos dividindo os dados em, por exemplo, 60% anticorpos, 20% supressão e 20% antígenos;
  - ◊ Para cada anticorpo, inicializar sua variável *contador* = 0;
  - ◊ Encontrar o anticorpo mais próximo a cada antígeno supressor;
  - ◊ Para cada antígeno reconhecido pelo anticorpo, contador++;
  - ◊ Eliminar os anticorpos onde *contador* = 0;
  - ◊ Gerar o novo arquivo de treino a partir dos anticorpos onde *contador* ≠ 0;
  - ◊ Fornecer este arquivo como entrada para treino de um algoritmo classificador.
- 

Neste trabalho, o mecanismo supressor gerou dados de treinamento para os classificadores 1NN, 10NN [34], discriminante linear [63, 34], árvore de decisão [63, 59, 34], máquina de vetor de suporte (SVM) [11], classificador *bayesiano* [63, 34] e *multilayer perceptron* (MLP) [35, 63, 59, 34].

Outro propósito desta nova abordagem é tentar evitar certos processos, como a poda em árvores de decisão, predições erradas no KNN e classificadores lineares além de se tentar lidar com o problema de desbalanceamento de classes.

Um ponto importante a se observar é que acredita-se que o mecanismo supressor não produza bons resultados em certos tipos de classificadores que melhoram seu desempenho com o aumento do conjunto de treinamento como, por exemplo, o classificador *bayesiano*.

Em todos os problemas de classificação, a determinação de um bom conjunto de treinamento é crucial para o modelo preditor final. O mecanismo proposto é sensível à forma como os dados da base estão distribuídos. Para reduzir esta sensibilidade e demonstrar que a escolha aleatória de um conjunto qualquer de treinamento não produzirá resultados tão bons quanto os conseguidos com a supressão, a técnica de *bagging* foi adotada.

Em seguida, para se tentar lidar com o problema de desbalanceamento, foi implementado junto com o mecanismo supressor um gerador de dados sintéticos – SMOTE – visto no capítulo 4, com o objetivo de igualar as quantidades das classes de cada base de dados. Os resultados para a classificação com o método seletor com *bagging* e SMOTE serão vistos nas próximas seções.

Para os estudos realizados, foram implementados em ambiente Matlab/C++ os seguintes classificadores:

1. KNN - Método dos k vizinhos mais próximos com distância Euclidiana [34];

2. Discriminate linear [63, 34];
3. Árvore de decisão - J4.8 [63, 59, 34];
4. SVM - Máquina de vetor de suporte com *kernel* de função de base radial [11];
5. Classificador *Bayesiano* [63, 34];
6. MLP (*multilayer perceptron*) com 3 camadas utilizando *backpropagation* [35, 63, 59, 34].

## 6.4 Resultados com *Bagging*

Nesta seção são expostos os resultados da supressão juntamente com o *bagging*. Foram selecionadas bases com características diferentes com relação ao tamanho, balanceamento, número de atributos e classes, de forma que o mecanismo pudesse ser testado em uma faixa abrangente de tipos de problemas. Estas bases foram descritas no capítulo 5. As tabelas a seguir mostram os valores das métricas definidas no capítulo 4 para diversos tipos de classificadores.

A Tabela 6.1 contém os primeiros resultados do mecanismo supressor hibridizado ao classificador 1NN, utilizando *bagging* com um comitê de 30 classificadores, ou 30 pacotes de inicialização, para 12 bases de dados diferentes. São mostrados os resultados para classificações utilizando todo o conjunto de treinamento produzido pelo *bootstrap* e, na segunda linha referente a cada base de dados – cujo nome é seguido da letra (S) – estão os resultados para os pacotes do *bagging* suprimidos.

As métricas utilizadas para avaliação do classificador foram: acurácia (*acc*), acurácia balanceada (*accB*), sensibilidade (*sens*), especificidade (*spec*), *F-measure* (*FMea*), *G-mean* (*GSP*), *G-mean* (*GSS*), *AUC* (*Area Under the ROC Curve*). O campo *Ant* representa o número de anticorpos utilizados na classificação. Quando não há supressão, *Ant* contém todo o conjunto de treino. Nos casos onde há supressão, *Ant* é a média dos conjuntos de treino do comitê de classificadores do *bagging*.

É possível ver pela Tabela 6.1 que os resultados foram bastante satisfatórios para o primeiro classificador referido, comparando-se a classificação com todo o conjunto de treino e com o conjunto suprimido. Somente as bases mais desbalanceadas, que são *ionosphere*, *glass*, *yeast* e *abalone* apresentaram resultados ruins com a supressão.

Observa-se, em particular, que a supressão da base de dados *abalone* foi bastante grande, reduzindo, em média, o conjunto de treinamento para apenas um exemplo. Isto certamente produzirá resultados ruins em todos os demais classificadores. Desta forma, a base não foi considerada nos outros experimentos. Faz-se necessário a sua investigação como um caso particular, visto que esta base é muito desbalanceada e com um grande número de registros. Possui, também, várias classes com apenas um ou dois exemplos, o

que impede o sucesso do método supressor, porque caso o exemplo único caia no conjunto de anticorpos, será suprimido; se cair no conjunto de antígenos, será visto como uma classificação incorreta. O mesmo acontece com as bases *Glass* e *Yeast*, onde o número de exemplos suprimidos ficou menor que o número de classes.

Analisando este fato dentro da metáfora imunológica, seria como se um patógeno invadisse o organismo e se mantivesse lá sem causar danos ou se reproduzir. Se o SI não o identifica como estranho e se este patógeno não é danoso, fica sem sentido a necessidade do disparo de mecanismos de defesa para o seu combate. Se o dado for um anticorpo, não terá função de defesa e, pela auto-regulação, acabará sendo eliminado.

Padrões de comportamento semelhantes podem ser vistos nas tabelas com os resultados para os outros classificadores. A Tabela 6.2 apresenta os resultados para o classificador 10NN. Neste caso, além das bases que apresentaram problemas com o classificador 1NN, para a *Wine* também não se produziram bons resultados.

O discriminante linear, que pode ser visto na tabela 6.3, no geral, apresentou bons resultados de classificação com a supressão. O mesmo acontece para árvore de decisão (Tabela 6.4), SVM (Tabela 6.5) e MLP (Tabela 6.6).

Conforme previsto anteriormente, os piores resultados foram para o classificador *bayesiano* (Tabela 6.7, visto que é sua característica melhorar o desempenho da predição à medida que lhe são apresentados mais exemplos de treinamento).

A partir da análise dos resultados obtidos, chegou-se à conclusão que o método supressor se mostra deficiente quando lida com bases contendo dados desbalanceados. Para tentar contornar este problema, experimentos foram realizados praticando-se a supressão sobre as bases balanceadas pelo algoritmo SMOTE. Como foi dito no capítulo 4, a função do SMOTE é balancear a base de dados por meio da geração de dados sintéticos. Acredita-se que com o balanceamento e a supressão, os resultados sejam melhorados e se consiga chegar a conjuntos de treinamento representativos para todas as classes do problema. Os experimentos realizados podem ser vistos na próxima seção.

Tabela 6.1: Resultados para o classificador 1NN sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.96	0.96	0.95	0.98	0.99	0.48	0.97	0.96	0.96	419
Câncer (S)	0.96	0.96	0.95	0.98	0.99	0.48	0.97	0.96	0.96	32
Diabetes	0.75	0.72	0.67	0.78	0.48	0.28	0.57	0.72	0.72	460
Diabetes (S)	0.78	0.78	0.79	0.78	0.44	0.28	0.59	0.78	0.78	79
<i>Monks-2</i>	0.96	0.92	0.88	0.96	0.64	0.37	0.75	0.92	0.92	360
<i>Monks-2</i> (S)	0.91	0.45	0.00	0.91	0.00	0.00	0.00	0.00	0.45	13
<i>Monks-3</i>	0.81	0.84	0.74	0.95	0.96	0.42	0.84	0.84	0.84	332
<i>Monks-3</i> (S)	0.85	0.87	0.78	0.95	0.96	0.43	0.87	0.86	0.87	47
<i>Ionosphere</i>	0.87	0.87	0.88	0.86	0.94	0.45	0.91	0.87	0.87	210
<i>Ionosphere</i> (S)	0.68	0.84	0.67	1.00	1.00	0.40	0.82	0.82	0.84	12
Íris	0.90	0.93	0.90	0.96	0.91	0.45	0.90	0.93	0.93	90
Íris (S)	0.93	0.96	0.93	0.98	0.95	0.47	0.94	0.96	0.96	9
<i>Well</i>	0.83	0.87	0.83	0.90	0.84	0.42	0.84	0.87	0.87	903
<i>Well</i> (S)	0.75	0.80	0.75	0.85	0.76	0.38	0.75	0.80	0.80	58
Contraceptivos	0.41	0.47	0.41	0.54	0.41	0.20	0.41	0.46	0.47	883
Contraceptivos (S)	0.44	0.51	0.44	0.58	0.44	0.22	0.44	0.50	0.51	83
<i>Wine</i>	0.97	0.98	0.97	0.99	0.97	0.49	0.97	0.98	0.98	106
<i>Wine</i> (S)	0.92	0.94	0.92	0.97	0.94	0.46	0.92	0.94	0.94	12
<i>Glass</i>	0.70	0.78	0.70	0.86	0.73	0.35	0.70	0.76	0.78	128
<i>Glass</i> (S)	0.56	0.71	0.56	0.85	0.85	0.32	0.67	0.68	0.71	2
<i>Yeast</i>	0.53	0.66	0.53	0.80	0.54	0.26	0.53	0.64	0.66	890
<i>Yeast</i> (S)	0.34	0.47	0.34	0.61	0.45	0.19	0.38	0.45	0.47	4
Abalone	0.20	0.45	0.20	0.69	0.20	0.10	0.20	0.34	0.45	2506
Abalone (S)	0.19	0.09	0.19	0.00	1.00	0.16	0.43	0.00	0.09	1

Tabela 6.2: Resultados para o classificador 10NN sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.96	0.97	0.96	0.98	0.99	0.49	0.97	0.97	0.97	419
Câncer (S)	0.91	0.93	0.89	0.98	0.99	0.47	0.94	0.93	0.93	32
Diabetes	0.79	0.78	0.75	0.81	0.54	0.31	0.64	0.78	0.78	460
Diabetes (S)	0.79	0.80	0.81	0.78	0.44	0.29	0.60	0.80	0.80	79
<i>Monks-2</i>	0.93	0.83	0.71	0.95	0.45	0.28	0.57	0.82	0.83	360
<i>Monks-2</i> (S)	0.91	0.45	0.00	0.91	0.00	0.00	0.00	0.00	0.45	13
<i>Monks-3</i>	0.91	0.91	0.87	0.96	0.96	0.46	0.92	0.91	0.91	332
<i>Monks-3</i> (S)	0.85	0.88	0.76	1.00	1.00	0.43	0.87	0.87	0.88	47
<i>Ionosphere</i>	0.90	0.92	0.88	0.95	0.98	0.46	0.93	0.92	0.92	210
<i>Ionosphere</i> (S)	0.66	0.33	0.66	0.00	1.00	0.40	0.81	0.00	0.33	12
Íris	0.93	0.96	0.93	0.98	0.95	0.47	0.94	0.96	0.96	90
Íris (S)	0.90	0.94	0.90	0.98	0.94	0.45	0.91	0.94	0.94	9
<i>Well</i>	0.77	0.82	0.77	0.86	0.78	0.39	0.77	0.82	0.82	903
<i>Well</i> (S)	0.74	0.79	0.74	0.85	0.75	0.37	0.74	0.79	0.79	58
Contraceptivos	0.45	0.51	0.45	0.58	0.45	0.22	0.45	0.50	0.51	883
Contraceptivos (S)	0.45	0.51	0.45	0.57	0.47	0.23	0.46	0.50	0.51	83
<i>Wine</i>	0.92	0.94	0.92	0.97	0.94	0.46	0.92	0.94	0.94	106
<i>Wine</i> (S)	0.39	0.68	0.39	0.97	0.95	0.25	0.58	0.60	0.68	12
<i>Glass</i>	0.67	0.77	0.67	0.86	0.77	0.36	0.72	0.75	0.77	128
<i>Glass</i> (S)	0.37	0.66	0.37	0.94	0.89	0.26	0.57	0.59	0.66	2
<i>Yeast</i>	0.54	0.68	0.54	0.82	0.59	0.28	0.56	0.66	0.68	890
<i>Yeast</i> (S)	0.29	0.15	0.29	0.00	1.00	0.23	0.54	0.00	0.15	4

Tabela 6.3: Resultados para o classificador Discriminante Linear sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.94	0.95	0.93	0.98	0.99	0.48	0.96	0.95	0.95	419
Câncer (S)	0.94	0.95	0.92	0.98	0.99	0.48	0.95	0.95	0.95	32
Diabetes	0.81	0.81	0.82	0.81	0.54	0.33	0.66	0.81	0.81	460
Diabetes (S)	0.77	0.77	0.78	0.77	0.42	0.27	0.57	0.77	0.77	79
<i>Monks-2</i>	0.93	0.96	1.00	0.92	0.18	0.15	0.43	0.96	0.96	360
<i>Monks-2</i> (S)	0.91	0.45	0.00	0.91	0.00	0.00	0.00	0.00	0.45	13
<i>Monks-3</i>	0.78	0.80	0.73	0.86	0.89	0.40	0.81	0.79	0.80	332
<i>Monks-3</i> (S)	0.79	0.80	0.74	0.87	0.89	0.40	0.81	0.80	0.80	47
<i>Ionosphere</i>	0.87	0.90	0.85	0.94	0.98	0.46	0.91	0.90	0.90	210
<i>Ionosphere</i> (S)	0.73	0.86	0.71	1.00	1.00	0.42	0.84	0.84	0.86	12
Íris	0.80	0.86	0.80	0.91	0.80	0.40	0.80	0.85	0.86	90
Íris (S)	0.80	0.86	0.80	0.91	0.80	0.40	0.80	0.85	0.86	9
<i>Well</i>	0.75	0.79	0.75	0.84	0.75	0.37	0.75	0.79	0.79	903
<i>Well</i> (S)	0.72	0.77	0.72	0.82	0.72	0.36	0.72	0.77	0.77	58
Contraceptivos	0.52	0.58	0.52	0.64	0.56	0.27	0.54	0.57	0.58	883
Contraceptivos (S)	0.54	0.60	0.54	0.66	0.59	0.28	0.56	0.59	0.60	83
<i>Wine</i>	0.97	0.98	0.97	0.99	0.97	0.49	0.97	0.98	0.98	106
<i>Wine</i> (S)	0.92	0.93	0.92	0.95	0.92	0.46	0.92	0.93	0.93	12
<i>Glass</i>	0.58	0.68	0.58	0.79	0.69	0.31	0.63	0.67	0.68	128
<i>Glass</i> (S)	0.40	0.69	0.40	0.99	0.98	0.27	0.61	0.62	0.69	2
<i>Yeast</i>	0.54	0.67	0.54	0.80	0.62	0.28	0.57	0.65	0.67	890
<i>Yeast</i> (S)	0.30	0.48	0.30	0.66	0.88	0.22	0.51	0.45	0.48	4

Tabela 6.4: Resultados para o classificador Árvore de Decisão sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.96	0.97	0.96	0.98	0.99	0.49	0.97	0.97	0.97	419
Câncer (S)	0.94	0.95	0.94	0.96	0.98	0.48	0.96	0.95	0.95	32
Diabetes	0.77	0.74	0.67	0.81	0.58	0.31	0.63	0.74	0.74	460
Diabetes (S)	0.77	0.74	0.71	0.78	0.48	0.29	0.58	0.74	0.74	79
<i>Monks-2</i>	0.93	0.84	0.75	0.93	0.27	0.20	0.45	0.84	0.84	360
<i>Monks-2</i> (S)	0.90	0.80	0.72	0.90	0.27	0.20	0.40	0.82	0.80	13
<i>Monks-3</i>	0.98	0.98	0.98	0.98	0.98	0.49	0.98	0.98	0.98	332
<i>Monks-3</i> (S)	0.97	0.97	0.98	0.96	0.96	0.49	0.97	0.97	0.97	47
<i>Ionosphere</i>	0.96	0.97	0.94	1.00	1.00	0.48	0.97	0.97	0.97	210
<i>Ionosphere</i> (S)	0.83	0.90	0.80	1.00	1.00	0.44	0.89	0.89	0.90	12
Íris	0.90	0.94	0.90	0.98	0.94	0.45	0.91	0.94	0.94	90
Íris (S)	0.43	0.69	0.43	0.95	0.92	0.26	0.60	0.62	0.69	9
<i>Well</i>	0.83	0.87	0.83	0.90	0.84	0.42	0.84	0.87	0.87	903
<i>Well</i> (S)	0.77	0.82	0.77	0.86	0.77	0.39	0.77	0.81	0.82	58
Contraceptivos	0.51	0.57	0.51	0.63	0.52	0.25	0.51	0.56	0.57	883
Contraceptivos (S)	0.55	0.61	0.55	0.67	0.58	0.28	0.56	0.61	0.61	83
<i>Wine</i>	0.97	0.98	0.97	0.99	0.97	0.49	0.97	0.98	0.98	106
<i>Wine</i> (S)	0.97	0.98	0.97	0.99	0.97	0.49	0.97	0.98	0.98	12
<i>Glass</i>	0.74	0.82	0.74	0.89	0.81	0.38	0.77	0.81	0.82	128
<i>Glass</i> (S)	0.35	0.17	0.35	0.00	1.00	0.26	0.59	0.00	0.17	2
<i>Yeast</i>	0.62	0.73	0.62	0.85	0.65	0.32	0.63	0.72	0.73	890
<i>Yeast</i> (S)	0.03	0.02	0.03	0.00	1.00	0.03	0.17	0.00	0.02	4

Tabela 6.5: Resultados para o classificador Máquina de Vetor de Suporte sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.95	0.94	0.97	0.92	0.96	0.48	0.96	0.94	0.94	419
Câncer (S)	0.98	0.98	0.98	0.98	0.99	0.49	0.98	0.98	0.98	32
Diabetes	0.80	0.79	0.77	0.81	0.54	0.32	0.65	0.79	0.79	460
Diabetes (S)	0.79	0.80	0.80	0.79	0.48	0.30	0.62	0.80	0.80	79
<i>Monks-2</i>	0.93	0.79	0.63	0.95	0.45	0.26	0.53	0.77	0.79	360
<i>Monks-2</i> (S)	0.91	0.45	0.00	0.91	0.00	0.00	0.00	0.00	0.45	13
<i>Monks-3</i>	0.96	0.96	0.95	0.98	0.98	0.48	0.96	0.96	0.96	332
<i>Monks-3</i> (S)	0.96	0.96	0.96	0.96	0.96	0.48	0.96	0.96	0.96	47
<i>Ionosphere</i>	0.93	0.93	0.94	0.91	0.96	0.47	0.95	0.93	0.93	210
<i>Ionosphere</i> (S)	0.66	0.33	0.66	0.00	1.00	0.40	0.81	0.00	0.33	12
Íris	0.93	0.96	0.93	0.98	0.95	0.47	0.94	0.96	0.96	90
Íris (S)	0.90	0.94	0.90	0.98	0.94	0.45	0.91	0.94	0.94	9
<i>Well</i>	0.75	0.80	0.75	0.85	0.77	0.38	0.76	0.80	0.80	903
<i>Well</i> (S)	0.75	0.80	0.75	0.85	0.75	0.37	0.75	0.79	0.80	58
Contraceptivos	0.53	0.59	0.53	0.64	0.62	0.28	0.56	0.58	0.59	883
Contraceptivos (S)	0.53	0.59	0.53	0.66	0.64	0.28	0.57	0.59	0.59	83
<i>Wine</i>	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00	1.00	106
<i>Wine</i> (S)	0.92	0.94	0.92	0.97	0.94	0.46	0.92	0.94	0.94	12
<i>Glass</i>	0.58	0.73	0.58	0.87	0.80	0.32	0.67	0.70	0.73	128
<i>Glass</i> (S)	0.47	0.62	0.47	0.77	0.76	0.28	0.59	0.60	0.62	2
<i>Yeast</i>	0.46	0.64	0.46	0.83	0.63	0.20	0.46	0.55	0.64	890
<i>Yeast</i> (S)	0.29	0.35	0.29	0.42	0.89	0.22	0.50	0.34	0.35	4

Tabela 6.6: Resultados para o classificador *Multi Layer Perceptron* sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.98	0.98	0.98	0.98	0.99	0.49	0.98	0.98	0.98	419
Câncer (S)	0.96	0.97	0.96	0.98	0.99	0.49	0.97	0.97	0.97	32
Diabetes	0.81	0.80	0.78	0.81	0.56	0.33	0.66	0.80	0.80	460
Diabetes (S)	0.79	0.77	0.74	0.81	0.56	0.32	0.64	0.77	0.77	79
<i>Monks-2</i>	0.93	0.83	0.71	0.95	0.45	0.28	0.57	0.82	0.83	360
<i>Monks-2</i> (S)	0.91	0.45	0.00	0.91	0.00	0.00	0.00	0.00	0.45	13
<i>Monks-3</i>	0.98	0.98	0.98	0.98	0.98	0.49	0.98	0.98	0.98	332
<i>Monks-3</i> (S)	0.95	0.95	0.93	0.96	0.96	0.47	0.95	0.95	0.95	47
<i>Ionosphere</i>	0.92	0.91	0.92	0.91	0.96	0.47	0.94	0.91	0.91	210
<i>Ionosphere</i> (S)	0.75	0.86	0.72	1.00	1.00	0.42	0.85	0.85	0.86	12
Íris	0.93	0.96	0.93	0.98	0.95	0.47	0.94	0.96	0.96	90
Íris (S)	0.90	0.93	0.90	0.96	0.91	0.45	0.90	0.93	0.93	9
<i>Well</i>	0.71	0.77	0.71	0.82	0.73	0.36	0.72	0.77	0.77	903
<i>Well</i> (S)	0.57	0.68	0.57	0.79	0.80	0.33	0.67	0.67	0.68	58
Contraceptivos	0.54	0.60	0.54	0.66	0.57	0.28	0.55	0.60	0.60	883
Contraceptivos (S)	0.52	0.57	0.52	0.63	0.55	0.26	0.53	0.56	0.57	83
<i>Wine</i>	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00	1.00	106
<i>Wine</i> (S)	0.92	0.93	0.92	0.95	0.92	0.46	0.92	0.93	0.93	12
<i>Glass</i>	0.51	0.64	0.51	0.77	0.70	0.29	0.59	0.63	0.64	128
<i>Glass</i> (S)	0.44	0.70	0.44	0.96	0.93	0.28	0.62	0.64	0.70	2
<i>Yeast</i>	0.57	0.70	0.57	0.82	0.63	0.29	0.59	0.68	0.70	890
<i>Yeast</i> (S)	0.29	0.40	0.29	0.52	0.85	0.21	0.49	0.38	0.40	4

Tabela 6.7: Resultados para o classificador *Bayesiano* sem supressão e com supressão utilizando *Bagging* (30 pacotes).

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Câncer	0.97	0.96	1.00	0.93	0.96	0.49	0.98	0.96	0.96	419
Câncer (S)	0.94	0.93	1.00	0.86	0.91	0.48	0.95	0.93	0.93	32
Diabetes	0.82	0.79	0.71	0.87	0.74	0.36	0.73	0.79	0.79	460
Diabetes (S)	0.79	0.76	0.66	0.87	0.74	0.35	0.70	0.76	0.76	79
<i>Monks-2</i>	0.91	0.75	0.50	1.00	1.00	0.33	0.71	0.71	0.75	360
<i>Monks-2</i> (S)	0.09	0.05	0.09	0.00	1.00	0.08	0.30	0.00	0.05	13
<i>Monks-3</i>	0.93	0.93	0.96	0.90	0.89	0.46	0.93	0.93	0.93	332
<i>Monks-3</i> (S)	0.90	0.90	0.87	0.94	0.95	0.45	0.91	0.90	0.90	47
<i>Ionosphere</i>	0.66	0.33	0.66	0.00	1.00	0.40	0.81	0.00	0.33	210
<i>Ionosphere</i> (S)	0.66	0.33	0.66	0.00	1.00	0.40	0.81	0.00	0.33	12
Íris	0.97	0.98	0.97	0.99	0.97	0.48	0.97	0.98	0.98	90
Íris (S)	0.03	0.06	0.03	0.08	0.03	0.01	0.03	0.00	0.06	9
<i>Well</i>	0.77	0.82	0.77	0.86	0.77	0.39	0.77	0.81	0.82	903
<i>Well</i> (S)	0.76	0.81	0.76	0.85	0.77	0.38	0.77	0.81	0.81	58
Contraceptivos	0.46	0.57	0.46	0.69	0.54	0.23	0.48	0.55	0.57	883
Contraceptivos (S)	0.45	0.51	0.45	0.56	0.57	0.25	0.51	0.49	0.51	83
<i>Yeast</i>	0.29	0.15	0.29	0.00	1.00	0.23	0.54	0.00	0.15	890
<i>Yeast</i> (S)	0.03	0.02	0.03	0.00	1.00	0.03	0.17	0.00	0.02	4

## 6.5 Resultados com *Bagging* e SMOTE

Nesta seção são apresentados os resultados do método seletor juntamente com o *bagging* e SMOTE. Foram selecionadas para experimentos as bases desbalanceadas *diabetes*, *ionosphere*, *Well*, *Glass* e *Yeast*. Ao contrário do que é apresentado no trabalho [9], nenhum mecanismo (exceto a supressão) para retirar o excesso de dados das bases foi adotado, pois acredita-se que a supressão serve também como uma forma de *undersampling*.

Nas tabelas de números 6.8 a 6.12 são mostradas as porcentagens de aumento para cada classe que serviram como entrada para o SMOTE. O número de vizinhos definido foi igual a 5, seguindo os experimentos realizados em [9]. As porcentagens foram definidas de forma que se chegasse a um total de exemplos com porcentagens de frequências parecido no total. Partiu-se da classe com maior número de instâncias e balancearam-se as demais a partir da maior.

Na Tabela 6.13 estão os resultados para o classificador 1NN. Comparando a classificação com o conjunto de treino completo e suprimido, vê-se que as métricas de avaliação dos classificadores ficaram muito parecidas. Com relação aos resultados somente com o *bagging* da Tabela 6.1, os resultados foram sempre melhores.

Para o classificador 10NN, os resultados (Tabela 6.14) foram melhores para a base *Glass*, comparados aos da Tabela 6.2. Os resultados para a base *Yeast* ficaram ruins.

Com o discriminante linear, na Tabela 6.15, os resultados foram sempre melhores ou semelhantes aos da Tabela 6.3, sem o SMOTE. O mesmo ocorre para árvore de decisão, SVM e MLP.

Já para o classificador *bayesiano* os resultados foram ruins, conforme previsto.

Tabela 6.8: Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados *Diabetes*.

Classe	Porcentagem
1	100
2	0

Tabela 6.9: Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados *Ionosphere*.

Classe	Porcentagem
1	0
2	300

Tabela 6.10: Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados *Well*.

Classe	Porcentagem
1	0
2	20
3	100

Tabela 6.11: Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados *Glass*.

Classe	Porcentagem
1	0
2	0
3	400
5	550
6	800
7	250

Tabela 6.12: Porcentagens de entrada para cada classe para o algoritmo SMOTE aplicado à base de dados *Yeast*.

Classe	Porcentagem
CYT	0
NUC	0
MIT	200
ME3	300
ME2	1000
ME1	1200
EXC	1500
VAC	3000
POX	6000
ERL	18000

## 6.6 Considerações Finais a Respeito do Mecanismo Seletor de Dados

Neste capítulo foi apresentado um novo método para seleção de exemplos de treinamento. A seleção é muito útil para diminuir custos, aumentar o desempenho computacional e eliminar exemplos pouco informativos. O método foi concebido para trabalhar junto a métodos de classificação, de maneira a melhorar o tempo de construção do classificador sem acarretar danos de acurácia. A abordagem teve inspiração no mecanismo de supressão do sistema imunológico.

Para realizar a supressão dos dados de treinamento, divide-se a base de dados original em três partes, anticorpos, supressão e antígenos. O reconhecimento é feito de acordo com o anticorpo mais próximo. Para cada antígeno da supressão reconhecido, produz-se

Tabela 6.13: Resultados para o classificador 1NN sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.84	0.85	0.80	0.90	0.93	0.43	0.86	0.85	0.85	621
Diabetes (S)	0.81	0.81	0.82	0.79	0.82	0.41	0.82	0.81	0.81	93
<i>Ionosphere</i>	0.97	0.97	0.96	0.98	0.96	0.48	0.96	0.97	0.97	437
<i>Ionosphere</i> (S)	0.94	0.92	0.85	0.99	0.98	0.45	0.91	0.92	0.92	78
<i>Well</i>	0.82	0.86	0.82	0.90	0.82	0.41	0.82	0.86	0.86	1098
<i>Well</i> (S)	0.76	0.81	0.76	0.86	0.76	0.38	0.76	0.81	0.81	107
<i>Glass</i>	0.76	0.86	0.76	0.95	0.76	0.38	0.76	0.83	0.86	286
<i>Glass</i> (S)	0.74	0.85	0.74	0.95	0.79	0.37	0.75	0.83	0.85	18
<i>Yeast</i>	0.84	0.91	0.84	0.98	0.86	0.42	0.85	0.91	0.91	4214
<i>Yeast</i> (S)	0.76	0.86	0.76	0.97	0.80	0.39	0.78	0.85	0.86	211

Tabela 6.14: Resultados para o classificador 10NN sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.80	0.84	0.75	0.92	0.96	0.42	0.85	0.83	0.84	621
Diabetes (S)	0.75	0.81	0.69	0.94	0.97	0.40	0.82	0.80	0.81	93
<i>Ionosphere</i>	0.97	0.96	0.92	1.00	1.00	0.48	0.96	0.96	0.96	437
<i>Ionosphere</i> (S)	0.84	0.83	0.66	1.00	1.00	0.40	0.81	0.81	0.83	78
<i>Well</i>	0.77	0.82	0.77	0.87	0.77	0.38	0.77	0.81	0.82	1098
<i>Well</i> (S)	0.73	0.80	0.73	0.86	0.77	0.37	0.74	0.79	0.80	107
<i>Glass</i>	0.71	0.83	0.71	0.95	0.78	0.35	0.72	0.80	0.83	286
<i>Glass</i> (S)	0.63	0.78	0.63	0.93	0.77	0.32	0.66	0.74	0.78	18
<i>Yeast</i>	0.11	0.50	0.11	0.89	0.01	0.01	0.04	0.00	0.50	4214
<i>Yeast</i> (S)	0.13	0.50	0.13	0.87	0.02	0.01	0.05	0.00	0.50	211

Tabela 6.15: Resultados para o classificador Discriminante Linear sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.74	0.74	0.80	0.68	0.68	0.37	0.74	0.74	0.74	621
Diabetes (S)	0.73	0.74	0.80	0.68	0.67	0.36	0.73	0.73	0.74	93
<i>Ionosphere</i>	0.90	0.90	0.88	0.91	0.80	0.42	0.84	0.90	0.90	437
<i>Ionosphere</i> (S)	0.90	0.88	0.82	0.94	0.87	0.42	0.84	0.88	0.88	78
<i>Well</i>	0.75	0.80	0.75	0.86	0.76	0.38	0.75	0.80	0.80	1098
<i>Well</i> (S)	0.74	0.79	0.74	0.85	0.74	0.37	0.74	0.79	0.79	107
<i>Glass</i>	0.77	0.87	0.77	0.98	0.90	0.40	0.82	0.86	0.87	286
<i>Glass</i> (S)	0.69	0.81	0.69	0.93	0.70	0.34	0.69	0.78	0.81	18
<i>Yeast</i>	0.62	0.78	0.62	0.95	0.72	0.32	0.65	0.75	0.78	4214
<i>Yeast</i> (S)	0.60	0.77	0.60	0.95	0.73	0.32	0.65	0.74	0.77	211

um estímulo de sobrevivência para o anticorpo. Sem nenhum estímulo, o anticorpo sai

Tabela 6.16: Resultados para o classificador Árvore de Decisão sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.77	0.77	0.80	0.74	0.77	0.39	0.78	0.77	0.77	621
Diabetes (S)	0.74	0.74	0.78	0.70	0.72	0.38	0.75	0.74	0.74	93
<i>Ionosphere</i>	0.95	0.94	0.93	0.95	0.89	0.46	0.91	0.94	0.94	437
<i>Ionosphere</i> (S)	0.91	0.91	0.92	0.91	0.78	0.42	0.85	0.91	0.91	78
<i>Well</i>	0.84	0.88	0.84	0.91	0.84	0.42	0.84	0.88	0.88	1098
<i>Well</i> (S)	0.77	0.82	0.77	0.87	0.77	0.38	0.77	0.82	0.82	107
<i>Glass</i>	0.82	0.90	0.82	0.97	0.84	0.41	0.83	0.89	0.90	286
<i>Glass</i> (S)	0.66	0.81	0.66	0.97	0.90	0.35	0.74	0.77	0.81	18
<i>Yeast</i>	0.82	0.90	0.82	0.98	0.84	0.41	0.83	0.89	0.90	4214
<i>Yeast</i> (S)	0.72	0.84	0.72	0.96	0.75	0.36	0.73	0.83	0.84	211

Tabela 6.17: Resultados para o classificador Máquina de Vetor de Suporte sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.78	0.78	0.80	0.76	0.79	0.40	0.80	0.78	0.78	621
Diabetes (S)	0.75	0.74	0.77	0.72	0.75	0.38	0.76	0.74	0.74	93
<i>Ionosphere</i>	0.97	0.96	0.96	0.97	0.93	0.47	0.95	0.96	0.96	437
<i>Ionosphere</i> (S)	0.96	0.96	0.95	0.96	0.91	0.47	0.93	0.96	0.96	78
<i>Well</i>	0.74	0.80	0.74	0.86	0.75	0.37	0.75	0.79	0.80	1098
<i>Well</i> (S)	0.74	0.80	0.74	0.86	0.76	0.37	0.75	0.79	0.80	107
<i>Glass</i>	0.68	0.82	0.68	0.97	0.85	0.35	0.73	0.77	0.82	286
<i>Glass</i> (S)	0.61	0.79	0.61	0.96	0.88	0.32	0.69	0.73	0.79	18
<i>Yeast</i>	0.61	0.78	0.61	0.96	0.79	0.29	0.64	0.72	0.78	4214
<i>Yeast</i> (S)	0.54	0.75	0.54	0.96	0.79	0.26	0.58	0.66	0.75	211

Tabela 6.18: Resultados para o classificador *Bayesiano* sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.69	0.69	0.75	0.63	0.63	0.34	0.69	0.69	0.69	621
Diabetes (S)	0.69	0.69	0.74	0.65	0.67	0.35	0.70	0.69	0.69	93
<i>Ionosphere</i>	0.32	0.16	0.32	0.00	1.00	0.24	0.56	0.00	0.16	437
<i>Ionosphere</i> (S)	0.32	0.16	0.32	0.00	1.00	0.24	0.56	0.00	0.16	78
<i>Well</i>	0.77	0.82	0.77	0.87	0.77	0.39	0.77	0.82	0.82	1098
<i>Well</i> (S)	0.76	0.81	0.76	0.86	0.76	0.38	0.76	0.81	0.81	107
<i>Glass</i>	0.68	0.82	0.68	0.96	0.84	0.37	0.75	0.80	0.82	286
<i>Glass</i> (S)	0.17	0.09	0.17	0.01	0.93	0.14	0.39	0.00	0.09	18
<i>Yeast</i>	0.07	0.03	0.07	0.00	1.00	0.06	0.26	0.00	0.03	4214
<i>Yeast</i> (S)	0.08	0.04	0.08	0.00	1.00	0.08	0.28	0.00	0.04	211

Tabela 6.19: Resultados para o classificador *Multi Layer Perceptron* sem supressão e com supressão utilizando *Bagging* (30 pacotes) e SMOTE.

Base de Dados	acc	accB	sens	spec	prec	FMea	GSP	GSS	AUC	Ant
Diabetes	0.74	0.73	0.77	0.70	0.73	0.37	0.75	0.73	0.73	621
Diabetes (S)	0.72	0.72	0.76	0.68	0.70	0.36	0.73	0.72	0.72	93
<i>Ionosphere</i>	0.97	0.96	0.94	0.98	0.96	0.47	0.95	0.96	0.96	437
<i>Ionosphere</i> (S)	0.92	0.90	0.81	0.98	0.96	0.44	0.88	0.89	0.90	93
<i>Well</i>	0.69	0.76	0.69	0.83	0.72	0.35	0.70	0.76	0.76	1098
<i>Well</i> (S)	0.70	0.77	0.70	0.84	0.74	0.36	0.72	0.77	0.77	107
<i>Glass</i>	0.57	0.75	0.57	0.92	0.71	0.30	0.62	0.69	0.75	286
<i>Glass</i> (S)	0.56	0.74	0.56	0.91	0.74	0.30	0.63	0.70	0.74	18
<i>Yeast</i>	0.63	0.79	0.63	0.95	0.71	0.33	0.66	0.77	0.79	4214
<i>Yeast</i> (S)	0.63	0.79	0.63	0.95	0.72	0.33	0.66	0.76	0.79	211

da população. Ao final do processo, o conjunto de treinamento resultante será composto pelos anticorpos sobreviventes.

O mecanismo supressor gerou dados de treinamento para os classificadores 1NN, 10NN, discriminante linear, árvore de decisão, máquina de vetor de suporte (SVM), classificador bayesiano e *multilayer perceptron* (MLP). Com a finalidade de reduzir esta sensibilidade dos classificadores, a técnica de *bagging* foi adotada. Para se tentar lidar com o problema de desbalanceamento, foi implementado junto com o mecanismo supressor um gerador de dados sintéticos – SMOTE.

Nos experimentos, foram selecionadas bases com características diferentes de forma que o mecanismo pudesse ser testado em uma faixa abrangente de tipos de problemas.

Os resultados com o *bagging* mostraram que o mecanismo de supressão é eficiente, adequado a vários tipos diferentes de classificadores – inclusive àqueles mais utilizados atualmente, como SVM e MLP. Há, no geral, uma redução grande dos dados de treinamento, sem que haja perdas significativas nas medidas de avaliação de desempenho. Há, também, ganho no tempo de processamento de todas as bases, como pode ser visto no exemplo da Tabela 6.20. Nela estão computados os tempos para processamento da base de dados *Yeast*. Esta base por conter um número grande de exemplos, demanda mais tempo de processamento. Principalmente em classificadores como KNN10 e SVM. Na tabela estão os classificadores e as unidades de tempo consumidas em média pelo processo de treinamento sem supressão e com os dados suprimidos, por pacote do *bagging*.

Em particular, houve casos que a supressão das bases de dados mais desbalanceadas produziu uma redução dos dados muito grande, acarretando em prejuízos no treinamento do classificador.

Para tentar amenizar este problema, experimentos foram realizados praticando-se a supressão sobre as bases balanceadas pelo algoritmo SMOTE. No geral, o desempenho da supressão gerou bons resultados e minimizaram-se os erros de seleção dos dados acar-

Tabela 6.20: Tempo gasto para treinamento dos classificadores.

Classificador	Unidades de tempo gastas
10NN	3
10NN (S)	2
Discriminante Linear	2
Discriminante Linear (S)	2
Árvore de Decisão	3
Árvore de Decisão (S)	2
SVM	3
SVM (S)	2
<i>Bayes</i>	3
<i>Bayes</i> (S)	2
MLP	3
MLP (S)	2

retados pelo balanceamento. Somente para o classificador *bayesiano* o algoritmo SeleSup produziu medidas insatisfatórias.

# Capítulo 7

## Considerações Finais

O sistema imunológico apresenta diversas características que lhe possibilitam detectar e combater um agente invasor no organismo de forma otimizada. Um dos mecanismos que o permitem ser capaz de defender o organismo contra todos os patógenos existentes – sem desperdício de energia e encontrando rapidamente os melhores agentes de defesa específicos, é a auto-regulação. O estudo dos fenômenos relacionados à auto-regulação e mecanismos de supressão serviram como inspiração para o desenvolvimento deste trabalho. Foram propostos, aqui, dois algoritmos imuno-inspirados. Um com o propósito de classificação e outro para seleção de dados de treinamento.

Na primeira proposta, foi introduzido um método classificador de dados determinístico da classe dos KNN, utilizando um mecanismo supressor dos dados de treinamento. Experimentos realizados utilizando alguns *benchmarks* contendo duas ou mais classes mostraram resultados satisfatórios para muitas bases. Contudo, este sistema apresenta alguns pontos negativos. Em primeiro lugar, seu caráter determinístico não permite maleabilidade no seu processo de aprendizado. Outro ponto a ser observado é que o sistema não detecta anomalias, como, por exemplo, casos raros que ocorrem dentro de certas classes. Logo, não se contempla a predição onde dados fora do padrão aparecem.

Uma aprimoração desta primeira idéia, já não mais objetivando-se classificar dados, mas obter uma base de treinamento suprimida que contenha os dados mais significativos dentro de todo o conjunto de treino, foi apresentada na segunda proposta. Outra meta foi a tentativa de amenizar problemas envolvendo a classificação de bases desbalanceadas e multi-classes. O método foi concebido para trabalhar junto a classificadores, de maneira a diminuir o tempo gasto com treinamento sem acarretar danos de acurácia. O mecanismo supressor gerou dados de treinamento para os classificadores 1NN, 10NN, discriminante linear, árvore de decisão, máquina de vetor de suporte (SVM), classificador *bayesiano* e *multilayer perceptron* (MLP). Juntamente com a supressão, técnicas de *bagging* e geração de dados sintéticos (SMOTE) foram adotadas.

Os resultados, em geral, mostraram que o mecanismo de supressão é eficiente, adequado a vários tipos diferentes de classificadores. Há uma redução grande dos dados

de treinamento, sem perdas significativas nas medidas de avaliação de desempenho. Por consequência, houve, também, ganho no tempo de processamento de todas as bases estudadas. Associado ao SMOTE, minimizaram-se os erros de seleção dos dados acarretados pelo balanceamento, exceto pelo classificador *bayesiano*.

Como continuação desta pesquisa, pretende-se aprimorar o algoritmo de seleção de dados. Uma primeira idéia, seria a inserção de dados sintéticos já na supressão, de forma a evitar a falta de exemplos de treino para as classes desbalanceadas.

Outra possibilidade é adaptar a escolha de dados para problemas de mineração de textos, onde o conjunto de treinamento é sempre muito grande. Acredita-se que, escolhendo textos mais representativos, pode-se conseguir meios mais eficientes para treinamento dos classificadores.

# Apêndice A

## Bases de Dados Utilizadas

Os experimentos foram realizados sobre doze bases de dados. Informações relevantes sobre estas bases estão resumidas na Tabela A.1. Como pode ser visto na tabela, as características de cada problema estudado variam com relação ao número de registros, proporções de classes e balanceamento. Oferecem, portanto, vários domínios de teste para a técnica proposta. As bases são as seguintes:

Base de Dados	Instâncias	Número de Atributos	Número de Classes
Câncer	699	9	2
Diabetes	768	8	2
<i>Monks-2</i>	601	6	2
<i>Monks-3</i>	554	6	2
<i>Ionosphere</i>	351	34	2
Íris	150	4	3
<i>Well</i>	1507	6	3
Contraceptivos	1473	9	3
<i>Wine</i>	178	13	3
<i>Glass</i>	214	9	6
<i>Yeast</i>	1484	8	10
Abalone	4177	7	28

Tabela A.1: Distribuição das Bases de Dados.

1. **Câncer** trata do diagnóstico do câncer de mama através da classificação do tumor como benigno ou maligno. O conjunto de dados é constituído de 9 entradas, 2 saídas e 699 exemplos. Todas as entradas são contínuas e 65,5% dos exemplos é formado por tumores benignos. Os dados foram obtidos com base no problema “Câncer de Mama de *Wisconsin*” do repositório de dados de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI) <sup>1</sup>. As porcentagens de cada classe podem ser vistas na Tabela A.2 e Figura A.1.

---

<sup>1</sup><http://kdd.ics.uci.edu>

Classes	Instâncias	Porcentagens
Maligno (0)	241	34.50%
Benigno (1)	458	65.50%

Tabela A.2: Porcentagens de cada classe referentes à base de dados Câncer.

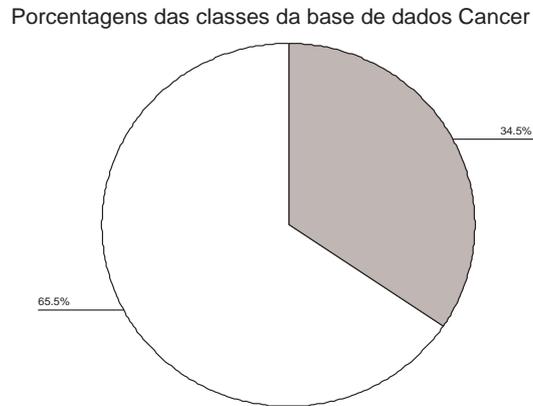


Figura A.1: Distribuição das classes da base de dados Câncer.

2. **Diabetes** trata do diagnóstico de diabetes dos índios Pima. As informações foram coletadas com base em dados pessoais (como idade e número de ocorrências de gravidez) e resultados de exames médicos (como por exemplo, pressão sanguínea, índice de massa corporal, resultado do teste de tolerância à glicose, etc). Através destes indicativos é possível determinar se o indivíduo é diabético ou não. O conjunto de dados é constituído de 8 entradas, 2 saídas e 768 exemplos. 65,1% dos exemplos são negativos. A Tabela A.3 e a Figura A.2 contêm as instâncias e porcentagens das duas classes.

Classes	Instâncias	Porcentagens
0	500	65.10%
1	268	34.90%

Tabela A.3: Porcentagens de cada classe referentes à base de dados Diabetes.

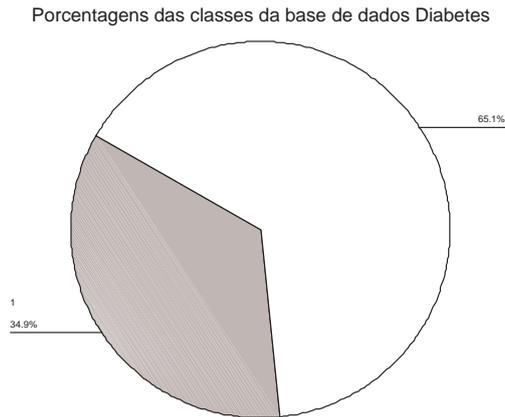


Figura A.2: Distribuição das classes da base de dados Diabetes.

3. **Monks-2** contém 6 atributos, 2 classes e 601 registros. Suas informações sobre as classes estão na Tabela A.4 e Figura A.3.

Classes	Instâncias	Porcentagens
0	395	65.72%
1	206	34.27%

Tabela A.4: Porcentagens de cada classe referentes à base de dados *Monks-2*.

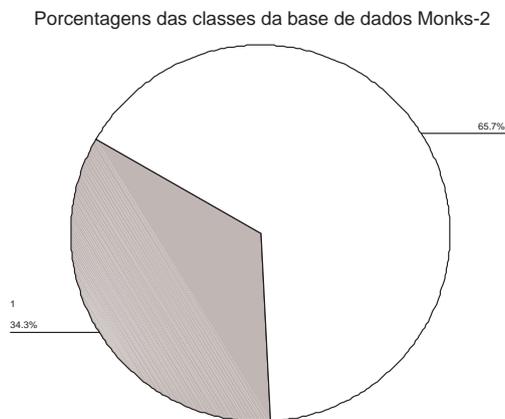


Figura A.3: Distribuição das classes da base de dados *Monks-2*.

4. **Monks-3** contém 6 atributos, 2 classes e 554 registros. As informações sobre as classes podem ser vistas na Tabela A.5 e Figura A.4. Esta base se diferencia da base anterior por apresentar 5% de dados ruidosos.
5. **Ionosphere** são dados de radar coletados por antenas que detectam elétrons livres na ionosfera. A base contém 351 instâncias, 34 atributos e 2 classes. A divisão de registros para cada classe é mostrada na Tabela A.6 e Figura A.5.

Classes	Instâncias	Porcentagens
0	266	48.02%
1	288	51.98%

Tabela A.5: Porcentagens de cada classe referentes à base de dados *Monks-3*.

Porcentagens das classes da base de dados Monks-3

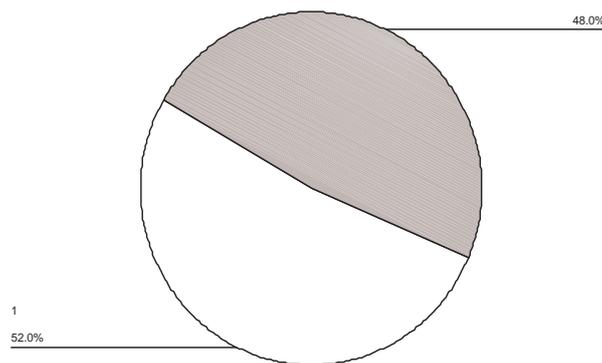


Figura A.4: Distribuição das classes da base de dados *Monks-3*.

Classes	Instâncias	Porcentagens
G	255	64.10%
B	96	35.90%

Tabela A.6: Porcentagens de cada classe referentes à base de dados *Ionosphere*.

Porcentagens das classes da base de dados Ionosphere

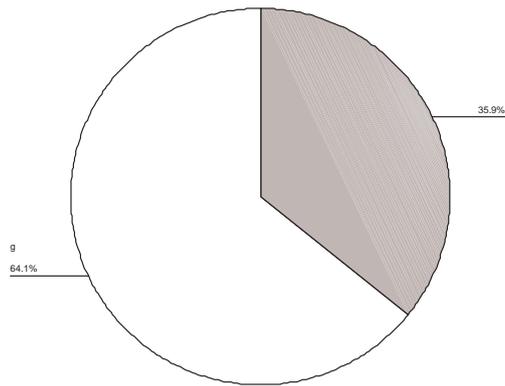


Figura A.5: Distribuição das classes da base de dados *Ionosphere*.

6. **Íris** contém 4 atributos e 3 classes com 50 instâncias cada, onde cada classe se refere a um tipo de planta íris. Uma classe é linearmente separável das demais. Os dados sobre esta base podem ser vistos na Tabela A.7 e Figura A.6.

Classes	Instâncias	Porcentagens
Iris Setosa	50	33.33%
Iris Versicolour	50	33.33%
Iris Virginica	50	33.33%

Tabela A.7: Porcentagens de cada classe referentes à base de dados *Íris*.

Porcentagens das classes da base de dados Iris

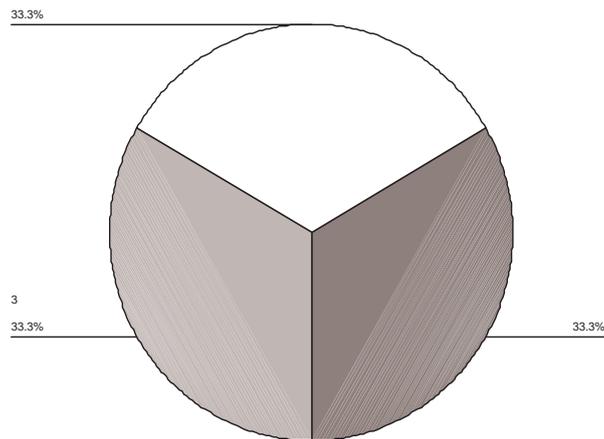


Figura A.6: Distribuição das classes da base de dados *Íris*.

7. **Well** contém 1506 registros, 6 atributos e 3 classes. As informações sobre esta base de dados estão na Tabela A.8 e Figura A.7.

Classes	Instâncias	Porcentagens
1	673	44.70%
2	509	33.80%
3	324	21.05%

Tabela A.8: Porcentagens de cada classe referentes à base de dados *Well*.

Porcentagens das classes da base de dados *Well*

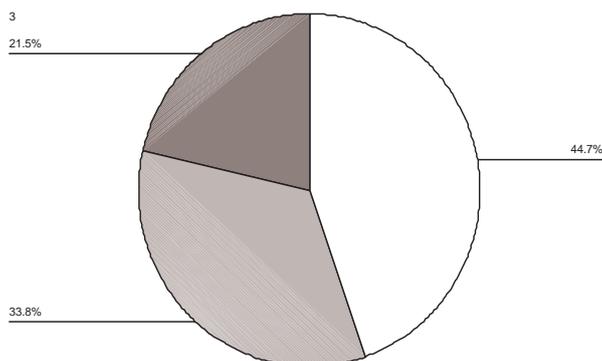


Figura A.7: Distribuição das classes da base de dados *Well*.

8. **Contraceptivos** é um subconjunto dos dados coletados em 1987 em uma pesquisa sobre anticoncepcionais na Indonésia. As amostras são sobre mulheres casadas que não estavam grávidas na época da entrevista. O objetivo era predizer o método anticoncepcional escolhido – podendo ser de utilização a longo ou curto prazo e sem uso – baseado em suas características demográficas e sócio-econômicas. Contém 1473 instâncias, 9 atributos e 3 classes. As informações sobre as classes estão na Tabela A.9 e figura A.8.

Classes	Instâncias	Porcentagens
1	629	42.70%
2	333	22.60%
3	511	34.70%

Tabela A.9: Porcentagens de cada classe referentes à base de dados *Contraceptivos*.

9. **Wine** contém dados resultantes de análises químicas de vinhos cultivados na mesma região da Itália, porém derivados de três vinhedos diferentes. As análises determinaram as quantidades de 13 constituintes encontrados em cada um dos 3 tipos de vinhos. A base contém 178 registros, como mostra a Tabela A.10 e a Figura A.9. Os dados nesta base são linearmente separáveis.

10. **Glass** envolve a classificação de tipos de vidro. Resultados de análises químicas

Porcentagens das classes da base de dados Contraceptivos

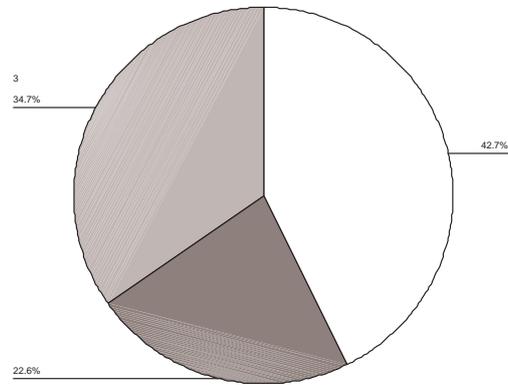


Figura A.8: Distribuição das classes da base de dados Contraceptivos.

Porcentagens das classes da base de dados Wine

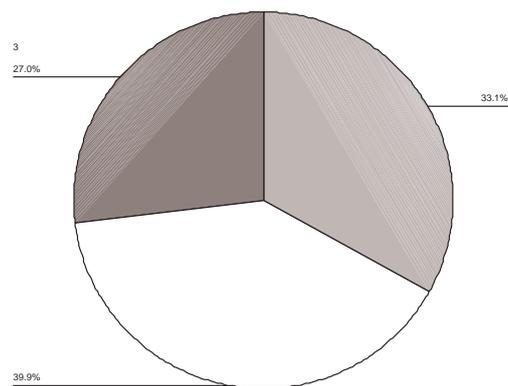


Figura A.9: Distribuição das classes da base de dados *Wine*.

Classes	Instâncias	Porcentagens
1	59	33.10%
2	71	39.90%
3	48	27.00%

Tabela A.10: Porcentagens de cada classe referentes à base de dados *Wine*.

mais o índice de refração são utilizados para classificar a amostra como janelas de edifícios processadas em flutuador ou não, janelas de veículos, containers, mesas ou lâmpadas. Esta classificação serve como auxílio à técnica forense para investigação criminal. A base contém 9 entradas, 6 saídas e 214 exemplos. As informações sobre esta base de dados estão na Tabela A.11 e Figura A.10. É possível ver, através dos dados contidos na tabela, que as classes estão desbalanceadas.

Classes	Instâncias	Porcentagens
1	70	32.71%
2	76	35.51%
3	17	7.94%
5	13	6.07%
6	9	4.20%
7	29	13.55%

Tabela A.11: Porcentagens de cada classe referentes à base de dados *Glass*.

Porcentagens das classes da base de dados *Glass*

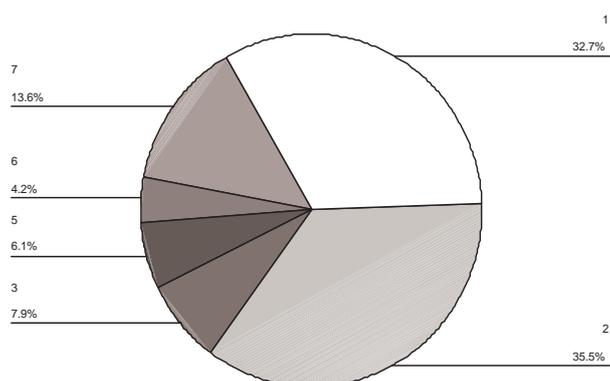


Figura A.10: Distribuição das classes da base de dados *Glass*.

- Yeast** é uma base que relaciona dados referentes à predição de sítios de localização de proteínas em células de bactérias. Contém 1484 registros, 8 atributos e 10 classes. A Tabela A.12 e a Figura A.11 contêm as informações sobre cada classe e mostra que este problema também apresenta desbalanceamento de classes.

Classes	Instâncias	Porcentagens
CYT	463	31.19%
NUC	429	28.9%
MIT	244	16.44%
ME3	163	10.98%
ME2	51	3.43%
ME1	44	2.96%
EXC	37	2.49%
VAC	30	2.02%
POX	20	1.34%
ERL	5	0.33%

Tabela A.12: Porcentagens de cada classe referentes à base de dados *Yeast*.

Porcentagens das classes da base de dados *Yeast*

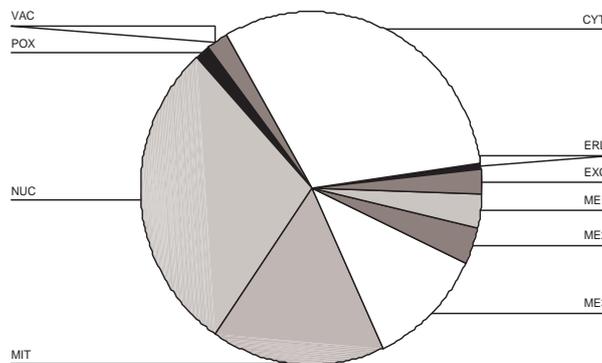


Figura A.11: Distribuição das classes da base de dados *Yeast*.

12. **Abalone** contém dados sobre a idade de um molusco da região da Califórnia *Abalone*, baseada em medições físicas da sua concha. Contém 4177 registros, 7 atributos (o atributo sexo foi retirado da base por ser nominal) e 28 classes. A distribuição das classes pode ser vista na Tabela A.13. A partir dos dados da tabela é possível concluir que esta é uma base de difícil classificação, visto que, além de haver desbalanceamento e uma distribuição muito grande de categorias, há classes onde o número de exemplos é pequeno para o aprendizado pelos classificadores.

Classes	Instâncias	Porcentagens
1	1	0.02%
2	1	0.02%
3	15	0.36%
4	57	1.36%
5	115	2.75%
6	259	6.20%
7	391	9.36
8	568	13.60%
9	689	16.49%
10	634	15.18%
11	487	11.66%
12	267	6.39%
13	203	4.86%
14	126	3.02%
15	103	2.46%
16	67	1.60%
17	58	1.39%
18	42	1.00%
19	32	0.77%
20	26	0.62%
21	14	0.33%
22	6	0.14%
23	9	0.21%
24	2	0.05%
25	1	0.02%
26	1	0.02%
27	2	0.05%
29	1	0.02%

Tabela A.13: Porcentagens de cada classe referentes à base de dados Abalone.

# Referências Bibliográficas

- [1] ABBAS A. K., LICHMAN A. H., P. J. S., *Imunologia celular e Molecular*. 2 ed. Rio de Janeiro, Revinter, 1998.
- [2] ADA, G. L., NOSSAL, G. J. V., “The Clonal Selection Theory”, *Scientific American*, v. 257, n. 2, pp. 50–57, 1987.
- [3] BAI, M., ZHAO, X., HOU, Z.-G., *et al.*, “Unified negative selection algorithm for anomaly detection”, In: *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, v. 1, pp. 4254–4258, 2006.
- [4] BALACHANDRAN, S., *Multi-shaped Detector generation using Real-valued representation for Anomaly Detection*, University of Memphis, 2005.
- [5] BLUM, A. L., LANGLEY, P., “Selection of relevant features and examples in machine learning”, *Artificial Intelligence*, v. 97, pp. 245–271, 1997.
- [6] BREIMAN, L., “Bagging Predictors”, *Machine Learning*, v. 24, n. 2, pp. 123–140, 1996.
- [7] BURNET, F. M., “A modification of Jerne’s Theory of Antibody Production Using the Concept of Clonal Selection”, *Australian Journal of Science*, v. 20, pp. 67–69, 1957.
- [8] CARTER, J. H., “The Immune System as a Model for Pattern Recognition and Classification”, *Journal of the American Medical Informatics Association*, v. 7, pp. 28–41, Jan/Feb 2000.
- [9] CHAWLA, N. V., BOWYER, K. W., KEGELMEYER, W. P., “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, v. 16, pp. 321–357, 2002.
- [10] CORMACK, D. H., *HAM Histologia*. 9 ed. Guanabara Koogan, 1991.
- [11] CRISTIANINI, N., SHAW-TAYLOR, J., *An introduction to support vector machines (and other kernel-based learning methods)*. Cambridge University Press, 2000.

- [12] DARWIN, C., *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859. 6th (final) ed'n 1872.
- [13] DASGUPTA, D., "Using immunological principles in anomaly detection", In: *Proceedings of the Artificial Neural Networks in Engineering (ANNIE96)*, St. Louis, MO, p. 443448, 1996.
- [14] DASGUPTA, D., YU, S., MAJUMDAR, N. S., "MILA - multilevel immune learning algorithm and its application to anomaly detection", *Soft Computing*, v. 9, n. 3, pp. 172–184, 2005.
- [15] DE CASTRO, L. N. & TIMMIS, J. I., "An artificial immune network for multimodal function optimization", In: *IEEE CEC'02*, pp. 699–674, 2002.
- [16] DE CASTRO, L. N., *Immune Engineering: Development of Computational Tools Inspired by the Artificial Immune Systems (in Portuguese)*. PhD thesis, DCA FEEC/UNICAMP, Campinas/SP, Brazil, August 1998 to May 2001.
- [17] DE CASTRO, L. N., TIMMIS, J., *Artificial Immune Systems: A New Computational Intelligence Approach*, v. 1. first ed. 2002.
- [18] DE CASTRO, L. N., ZUBEN, F. J. V., *aiNet: An Artificial Immune Network for Data Analysis*, ch. XII, pp. 231–259. Idea Group Publishing, USA, Chapter XII, 2001.
- [19] DE CASTRO, L. N., ZUBEN, F. J. V., "ainet: An artificial immune network for data analysis", In: *Data Mining: A Heuristic Approach* (HUSSEIN A. ABBASS, R. A. S., NEWTON, C. S., eds.), Idea Group Publishing, USA, 2001.
- [20] D.J. NEWMAN, S. HETTICH, C. B., MERZ, C., "UCI Repository of machine learning databases", 1998.
- [21] DONG, E. S., DONG, Y. G., JIA, H. C., *et al.*, "Application of a negative selection algorithm to anomaly detection for aircraft electrical power system", In: *Proceedings of the Third International Symposium on Instrumentation Science and Technology, Vol 3* (TAN, J., WEN, X., eds.), pp. 1161–1164, 2004.
- [22] ESPÍNDOLA, R. P., EBECKEN, N. F. F., "On extending f-measure and g-mean metrics to multi-class problems", In: *Sixth International Conference on Data Mining, Text Mining and their Business Applications*, v. 35, (Skiathos, Greece), pp. 25–34, Wessex Institute of Technology, UK, WIT Press, 25 - 27 May 2005.
- [23] FANG, L., LE-PING, L., "Unsupervised anomaly detection based on an evolutionary artificial immune network", In: *Applications of evolutionary computing: EvoWorkshops 2005 : EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoMUSART, and*

*EvoSTOC : Lausanne, Switzerland, March 30-April 1, 2005 : proceedings* (ROTHLAUF, F., ed.), (Lausanne, Switzerland), pp. 166–174, Berlin, 2005.

- [24] FIGUEREDO, G. P., DE CARVALHO, L. A. V., BARBOSA, H. J. C., “Coevolutionary genetic algorithms to simulate the immune system’s gene libraries evolution”, In: *Advances in Natural Computation: First International Conference, Lecture Notes in Computer Science, Changsha - China*, v. 3611, pp. 941–944, Springer Verlag - Berlin, 2005.
- [25] FIGUEREDO, G. P., EBECKEN, N. F. F., BARBOSA, H. J. C., “The supraic algorithm: A suppression immune based mechanism to find a representative training set in data classification tasks”, In: *ICARIS*, v. 4628 of *Lecture Notes in Computer Science*, pp. 59–70, Springer, 2007.
- [26] FORREST, S., SMITH, R. E., JAVORNIK, B., *et al.*, “Using Genetic Algorithms to Explore Pattern Recognition in the Immune System”, *Evolutionary Computation*, v. 1, n. 3, pp. 191–211, 1993.
- [27] FREITAS, A., TIMMIS, J., “Revisiting the Foundations of Artificial Immune Systems for Data Mining”, *IEEE Transactions on Evolutionary Computation*, v. 11, pp. 521–540, 2007.
- [28] GOLUB, E. S., “Is the function of the immune system only to protect?”, In: *Theoretical and Experimental Insights into Immunology*, v. 66 of *H: Cell Biology*, pp. 15–26, NATO ASI Series, 1992.
- [29] GONZALES, L. J., *A Self-Adaptive Evolutionary Negative Selection Approach for Anomaly Detection*. PhD thesis, Nova Southeastern University, 2005.
- [30] GOODMAN, D., BOGGESE, L., WATKINS, A., “An investigation into the source of power for airs, an artificial immune classification system”, In: *Intl. Joint Conference on Neural Networks (IJCNN’03)*, (Portland, OR, USA), pp. 1678–1683, Springer, July 2003.
- [31] GORDON, A., *Classification*. Chapman and Hall, London, 1981.
- [32] GREENSMITH, J., TWYXCROSS, J., AICKELIN, U., “Dendritic cells for anomaly detection”, In: *IEEE Congress on Evolutionary Computation; CEC*, pp. 664–671, 2006.
- [33] GUZELLA, T. S., MOTA-SANTOS, T. A., CAMINHAS, W. M., “Towards a novel immune inspired approach to temporal anomaly detection”, In: *ICARIS* (DE CASTRO, L. N., ZUBEN, F. J. V., KNIDEL, H., eds.), v. 4628 of *Lecture Notes in Computer Science*, pp. 119–130, Springer, 2007.

- [34] HAN, J., KAMBER, M., *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.
- [35] HAYKIN, S., *Redes neurais, princípios e práticas*. Bookman, 2001.
- [36] HIGHTOWER, R., FORREST, S., PERELSON, A. S., “The evolution of emergent organization in immune system gene libraries”, In: *Proceedings of the Sixth International Conference on Genetic Algorithms* (ESHELMAN, L., ed.), (San Francisco, CA), pp. 344–350, Morgan Kaufmann, 1995.
- [37] HOFMEYR, S. A., FORREST, S., “Architecture for an Artificial Immune System”, *Evolutionary Computation*, v. 8, n. 4, pp. 443–473, 2000.
- [38] JANEWAY, C. A., TRAVERS, P., WALPORT, M., *et al.*, *Immunobiologia: O sistema imune na saúde e na doença (In Portuguese)*. 5 ed. Artes Médicas, 2001.
- [39] JERNE, N. K., “Towards a network theory of the immune system”, *Ann. Immunol (Inst. Pasteur)*, v. 125C, n. 3, pp. 73–89, 1974.
- [40] JOHN, G. H., KOHAVI, R., PFLEGER, K., “Irrelevant features and the subset selection problem”, pp. 121–129, Morgan Kaufmann, 1994.
- [41] KARNIK, N., MENDEL, J., “Applications of type-2 fuzzy logic systems: handling the uncertainty associated with surveys”, In: *FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings*, v. 3, pp. 1546–51, 1999.
- [42] KEPLER, T. B., PERELSON, A. S., “Cyclic reentry of germinal center B cells and the efficiency of affinity maturation”, *Immunol. Today* 14, v. 14, pp. 412–415, 1993.
- [43] KEPLER, T. B., PERELSON, A. S., “Somatic hypermutation in B cells: An optimal control treatment.”, *J. Theoret. Biol.*, v. 164, pp. 37–64, 1993.
- [44] KIM, M., KIM, M., SEO, J., “Network anomaly behavior detection using an adaptive multiplex detector”, In: *Computational Science and Its Applications - Iccsa 2006, Pt 3* (GAVRILOVA, M., GERVASI, O., KUMAR, V., *et al.*, eds.), v. 3982 of *Lecture Notes in Computer Science*, pp. 154–162, 2006.
- [45] KLIR, G. J., YUAN, B., *Fuzzy sets and fuzzy logic: theory and applications*. Upper Saddle River, NJ, USA, Prentice-Hall, Inc., 1995.
- [46] KNIGHT, T., TIMMIS, J., “Aine: An immunological approach to data mining”, In: *IEEE International Conference on Data Mining* (CERCONE, N., LIN, T., WU, X., eds.), (San Jose, CA. USA), pp. 297–304, IEEE, December 2001.

- [47] KNIGHT, T., TIMMIS, J., “Assessing the performance of the resource limited artificial immune system AINE”, Tech. Rep. 3-01, Canterbury, Kent. CT2 7NF, May 2001.
- [48] LANG, F. H., LI, J., YANG, Y. X., “A novel fuzzy anomaly detection method based on clonal selection clustering algorithm”, In: *Machine Learning and Cybernetics; ICMLC* (YEUNG, D. S., LIU, Z. Q., WANG, X. Z., *et al.*, eds.), v. 3930 of *Lecture Notes in Artificial Intelligence*, pp. 642–651, 2006.
- [49] LITTLESTONE, N., MESTERHARM, C., “An apobaesysian relative of winnow”, In: *Advances in neural information processing systems*, v. 9, MIT Press, 1997.
- [50] LUO, H., FANG, B., YUN, X., “Anomaly detection in smtp traffic”, In: *Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on*, pp. 408–413, 2006.
- [51] MARWAH, G., BOGGESS, L., “Artificial Immune Systems for classification : Some issues”, 2002.
- [52] MITCHELL, T., *Machine Learning*. McGraw-Hill, 1997.
- [53] OPREA, M., PERELSON, A., “Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts”, *J. Immunol.*, v. 158, pp. 5155–5162, 1997.
- [54] OPREA, M., FORREST, S., “How the immune system generates diversity: Pathogen space coverage with random and evolved antibody libraries”, In: *Proceedings of the Genetic and Evolutionary Computation Conference* (BANZHAF, W., DAIDA, J., EIBEN, A. E., *et al.*, eds.), v. 2, (Orlando, Florida, USA), pp. 1651–1656, Morgan Kaufmann, 13-17 1999.
- [55] OPREA, M., KEPLER, T. B., “Genetic plasticity of V genes under somatic hypermutation: Statistical analyses using a new resampling-based methodology.”, *Genome Research*, v. 9 (12), pp. 1294–1304, 1999.
- [56] OSTASZEWSKI, M., SEREDYNSKI, F., BOUVRY, P., “Immune anomaly detection enhanced with evolutionary paradigms”, In: *Genetic and Evolutionary Computation Conference; GECCO*, (Seattle, Washington, USA), pp. 119–126, ACM Press, 2006.
- [57] PERELSON, A. S., OSTER, G. F., “The Shape Space Model”, *Journal of Theoretical Biology*, v. 81, pp. 645–670, 1979.
- [58] PERELSON, A. S., WEISBUCH, G., “Immunology for physicists”, *Reviews of Modern Physics*, v. 69, p. 1219, 1997.

- [59] QUINLAN, J. R., *C4.5: Programs for machine learning*. Morgan Kaufmann, 1993.
- [60] RON, J., “The evolution of secondary organization in immune system gene libraries”.
- [61] ROSS, T. J., *Fuzzy Logic with Engineering Applications*. New York, McGraw Hill, 1995.
- [62] RUMJANEK, V. M., “Próprio e estranho: é essa a questão?”, *Ciência Hoje*, v. 29, p. 40, 2001.
- [63] S. RUSSEL, P. N., *Artificial Intelligence: A Modern Approach*. Prentice Hall International, 1995.
- [64] SCHAPIRE, R. E., “The strength of weak learnability”, *Machine Learning*, v. 5, pp. 197–227, 1990.
- [65] SEREDYNSKI, F., BOUVRY, P., “Some issues in solving the anomaly detection problem using immunological approach”, In: *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International*, pp. 188b–188b, 2005.
- [66] SEREDYNSKI, F., BOUVRY, P., RUTKOWSKI, D. R., “Anomaly detection system for network security: Immunity-based approach”, In: *Intelligent Information Processing and Web Mining* (KLOPOTEK, M. A., WIERZCHON, S. T., TROJANOWSKI, K., eds.), *Advances in Soft Computing*, pp. 486–490, 2005.
- [67] SHIMURA, J., “Somatic mutations in immunoglobulin v gene determine the structure and function of the protein - an evidence from homology modeling”, In: *Bio-computing: Proceedings of the 1996 Pacific Symposium* (HUNTER, L., KLEIN, T., eds.), World Scientific Publishing, 1996.
- [68] SINGH, S. P. N., “Anomaly detection using negative selection based on the r-contiguous matching rule”, In: *International Conference on Artificial Immune Systems, ICARIS, 2002*.
- [69] STIBOR, T., *On the Appropriateness of Negative Selection for Anomaly Detection and Network Intrusion Detection*. PhD thesis, Darmstadt University of Technology, 2006.
- [70] STIBOR, T., MOHR, P., TIMMIS, J., *et al.*, “Is negative selection appropriate for anomaly detection?”, In: *Genetic and Evolutionary Computation Conference; GECCO*, (Washington DC, USA), pp. 321–328, ACM Press, 2005.

- [71] TIMMIS, J., EDMONDS, C., “A Comment on opt-AINet: An Immune Network Algorithm for Optimisation”, In: *Genetic and Evolutionary Computation* (ET AL, D. K., ed.), v. 3102 of *Lecture Notes in Computer Science*, pp. 308–317, Springer, July 2004.
- [72] TIMMIS, J., NEAL, M., “A resource limited artificial immune system for data analysis”, *Knowledge Based Systems*, v. 14, n. 3-4, pp. 121–130, 2001.
- [73] TIMMIS, J., NEAL, M., HUNT, J., “An Artificial Immune System for Data Analysis”, *Biosystems*, v. 55, n. 1/3, pp. 143–150, 2000.
- [74] TIMMIS, J., NEAL, M. J., “A Resource Limited Artificial Immune System for Data Analysis”, *Research and Development in Intelligent Systems XVII*, pp. 19–32, December 2000. Proceedings of ES2000, Cambridge, UK.
- [75] TIMMIS, J., KNIGHT, T., “Artificial Immune Systems: Using the Immune System as Inspiration for Data Mining”, In: *Data Mining: A Heuristic Approach* (ABBASS, H. A., SARKER, R. A., NEWTON, C. S., eds.), ch. XI, pp. 209–230, Group Idea Publishing, September 2001.
- [76] TIMMIS, J., NEAL, M., “A resource limited artificial immune system for data analysis”, *Knowledge Based Systems*, v. 14, pp. 121–130, June 2001.
- [77] TIZARD, I., *Introdução à Imunologia Veterinária*. 2 ed. 1985.
- [78] VAZ, N. M., “A teoria burnetiana hoje”, *Ciência Hoje*, v. 29, p. 40, 2001.
- [79] VORONOI, G., “Nouvelles applications des paramètres continus à la théorie des formes quadratiques”, *J. für die Reine und Angewandte Mathematik*, v. 133, pp. 97–178, 1907.
- [80] WATKINS, A., *AIRS: A RESOURCE LIMITED IMMUNE CLASSIFIER*, Mississippi State University, 2001.
- [81] WATKINS, A., *Exploiting Immunological Metaphors in the Development of Serial, Parallel, and Distributed Learning Algorithms*. PhD thesis, University of Kent, Canterbury, UK, March 2005.
- [82] WATKINS, A., TIMMIS, J., “Artificial immune recognition system AIRS: Revisions and refinements”, In: *1st Intl. Conference on Artificial Immune Systems (ICAIRIS2002)* (TIMMIS, J., BENTLEY, P., eds.), (University of Kent), pp. 173–181, September 2002.

- [83] WATKINS, A., TIMMIS, J., “Exploiting parallelism inherent in AIRS, an artificial immune classifier”, In: *In Proc. of the 3rd Intl. Conference on Artificial Immune Systems (ICARIS 2004)* (NICOSIA, G., CUTELLO, V., BENTLEY, P., *et al.*, eds.), (Catania, Italy), pp. 427–438, 2004.
- [84] WATKINS, A., TIMMIS, J., BOGGESS, L., *Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm*, v. 5, pp. 291 – 317. Springer Netherlands, September 2004.
- [85] WATKINS, A. B., BOGGESS, L. C., “A new classifier based on resource limited artificial immune systems”, In: *Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence, Honolulu, HI, USA*, pp. 1546–1551, May 12-17 2002.
- [86] WATKINS, A. B., BOGGESS, L. C., “A resource limited artificial immune classifier”, In: *Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence, Honolulu, HI, USA*, pp. 926–931, May 12-17 2002.
- [87] WEISS, G. M., “Mining with rarity: a unifying framework”, *SIGKDD Explor. Newsl.*, v. 6, n. 1, pp. 7–19, 2004.
- [88] XU, J., YOU, J., LIU, F., “A fuzzy rules based approach for performance anomaly detection”, In: *Networking, Sensing and Control, 2005. Proceedings. 2005 IEEE*, pp. 44–48, 2005.
- [89] ZHANG, F. B., YANG, Y. T., WANG, S. W., “Anomaly detection using real-valued negative selection”, In: *Istm/2005: 6th International Symposium on Test and Measurement, Vols 1-9, Conference Proceedings* (WEN, T. D., ed.), pp. 949–952, 2005.
- [90] ZHANG, Y., “A novel immune detection algorithm for anomaly detection”, In: *Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation*, pp. 1441–1446, 2005.

# Livros Grátis

( <http://www.livrosgratis.com.br> )

Milhares de Livros para Download:

[Baixar livros de Administração](#)

[Baixar livros de Agronomia](#)

[Baixar livros de Arquitetura](#)

[Baixar livros de Artes](#)

[Baixar livros de Astronomia](#)

[Baixar livros de Biologia Geral](#)

[Baixar livros de Ciência da Computação](#)

[Baixar livros de Ciência da Informação](#)

[Baixar livros de Ciência Política](#)

[Baixar livros de Ciências da Saúde](#)

[Baixar livros de Comunicação](#)

[Baixar livros do Conselho Nacional de Educação - CNE](#)

[Baixar livros de Defesa civil](#)

[Baixar livros de Direito](#)

[Baixar livros de Direitos humanos](#)

[Baixar livros de Economia](#)

[Baixar livros de Economia Doméstica](#)

[Baixar livros de Educação](#)

[Baixar livros de Educação - Trânsito](#)

[Baixar livros de Educação Física](#)

[Baixar livros de Engenharia Aeroespacial](#)

[Baixar livros de Farmácia](#)

[Baixar livros de Filosofia](#)

[Baixar livros de Física](#)

[Baixar livros de Geociências](#)

[Baixar livros de Geografia](#)

[Baixar livros de História](#)

[Baixar livros de Línguas](#)

[Baixar livros de Literatura](#)  
[Baixar livros de Literatura de Cordel](#)  
[Baixar livros de Literatura Infantil](#)  
[Baixar livros de Matemática](#)  
[Baixar livros de Medicina](#)  
[Baixar livros de Medicina Veterinária](#)  
[Baixar livros de Meio Ambiente](#)  
[Baixar livros de Meteorologia](#)  
[Baixar Monografias e TCC](#)  
[Baixar livros Multidisciplinar](#)  
[Baixar livros de Música](#)  
[Baixar livros de Psicologia](#)  
[Baixar livros de Química](#)  
[Baixar livros de Saúde Coletiva](#)  
[Baixar livros de Serviço Social](#)  
[Baixar livros de Sociologia](#)  
[Baixar livros de Teologia](#)  
[Baixar livros de Trabalho](#)  
[Baixar livros de Turismo](#)