Universidade Federal de Juiz de Fora
Faculdade de Engenharia
Bacharelado em Engenharia Computacional

SAINET com raio adaptativo: estimativa do raio mínimo e modificação no cálculo da densidade

Ana Lívia Soares Silva de Almeida

JUIZ DE FORA
DEZEMBRO, 2018

SAINET com raio adaptativo: estimativa do raio mínimo e modificação no cálculo da densidade

Ana Lívia Soares Silva de Almeida

Universidade Federal de Juiz de Fora

Faculdade de Engenharia

Departamento de Mecânica Aplicada e Computacional

Bacharelado em Engenharia Computacional

Orientador: Luciana Conceição Dias Campos

Coorientador: Heder Soares Bernardino

JUIZ DE FORA

DEZEMBRO, 2018

SAINET com raio adaptativo: estimativa do raio mínimo e modificação no cálculo da densidade

Ana Lívia Soares Silva de Almeida

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO FACULDADE DE ENGENHARIA DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM ENGENHARIA COMPUTACIONAL.

Aprovada por:

Luciana Conceição Dias Campos
Doutora

Heder Soares Bernardino
Doutor

Leonardo Goliatt da Fonseca
Doutor

Stênio Sã Rosário Furtado Soares
Doutor

JUIZ DE FORA

03 DE DEZEMBRO, 2018

Às mãos divinas que até aqui me guiaram, a minha inifinita gratidão.

À minha mãe, Rosely, por tudo que fez por mim até hoje para que eu me tornasse quem hoje sou.

Ao meu avô Oswaldo (in memorian) pela semente plantada na infância, pelo mundo que me apresentou, pelo incentivo à minha curiosidade, e pela saudade que não diminui com anos.

Resumo

Há diversos algoritmos baseados nas mais diferentes abordagens para tarefas de agrupamento e classificação de conjuntos de dados na literatura. Dentre os algoritmos baseados em densidade, alguns funcionam posicionando protótipos para representar os dados da base reduzindo assim o volume de dados manipulado. Um dos primeiros algoritmos imuno-inspirados com esta abordagem, intitulado *Artificial Immune Network* (aiNet), é uma rede imunológica artificial que tem como finalidade resolver problemas de agrupamento de dados. A partir da aiNet, vários outros algoritmos foram propostos para ser aplicados em outros tipos de problemas. Uma versão supervisonada da aiNet foi proposta para aplicação em problemas de classificação, chamada SAINET (AINET Supervisionada). Entretanto, o algoritmo da aiNet não se mostrava eficiente para algumas distribuições de dados e um outro algortimo com abordagem similar foi proposto: *Adaptative Radius Immune Algorithm* (ARIA) objetivando melhorar o resultado do agrupamento. Com base nisto, foi proposta uma modificação no algortimo original da SAINET que pretendia avaliar o resultado da classificação quando substituindo-se a aiNet pela ARIA para construção da população de protótipos, sendo portanto uma versão da SAINET com raio adaptativo. A literatura apresenta alguns estudos a cerca da eficiência do algoritmo da ARIA no que diz respeito à preservação da distribuição de densidade dos dados ao posicionar os protótipos para representá-los, uma vez que a ARIA não é capaz de preservar a densidade dos dados em grandes dimensões. Contudo, não há muita informação sobre como os parâmetros da ARIA que não são ajustados pelo algoritmo podem ser definidos pelo usuário. Este trabalho se destina a propor uma maneira de ajustar o raio mínimo dos anticorpos, parâmetro da ARIA que é fundamental no posicionamento dos protótipos e não é claramente determinável, além de verificar se a alteração no algoritmo para preservar a densidade dos dados tem algum efeito no resultado da classificação realizada pela SAINET com raio adaptativo.

Palavras-chave: Redes Imunológicas Artificiais, classificação, raio adaptativo, SAINET,

ARIA.

Abstract

There are several algorithms based on the most different approaches to perform data clustering and data classification in literature. Among those based on density, some are designed to generate prototypes in order to represent the dataset reducing the volume of data managed. One of the first of these immune-inspired is the *Artificial Immune Network* (aiNet), which is an artificial immune network designed to perform data clustering. From aiNet, many other algorithms have been proposed to be applied to other kind of problems. A supervised version of aiNet was porposed to perform data classification, called SAINET (supervised AINET). However, aiNet did not present satisfactory performance when applied to certain data distributions and another network was presented: the *Artifical Radius Immune Algorith* (ARIA) aiming to achieve better results where aiNet has not seem to be efficient. Based on that, a modification on SAINET original algorithm was proposed by replacing aiNet impementation by ARIA on building the network population, being though a SAINET wiht adaptative radius. Literature presents several studies about ARIA's efficiency regarding density preservation. Nevertheless, not much has been said about how to set its parameters that are not adjusted by the algorithm. The present work proposes a possibility to adjust ARIA's antibodies minimum radius, a parameter that is set by the user which is extremely important to network's performance and is not clear how to do. Further, we inted to verify whether the modification proposed to ARIA in order to preserve density affects SAINET classification performance.

Keywords: Artificial Immune Network, classification, adaptative radius, SAINET, ARIA.

Agradecimentos

Agradeço a todos que me apoiaram e incentivaram de alguma forma até aqui.

A Deus, em todas as suas formas de se fazer presente e me conduzir.

À minha mãe, Rosely, por todo amor e dedicação, por sempre acreditar em mim e jamais desisitir. Ao meu avô Oswaldo, quem me incentivou o gosto pela ciência e a quem eu prometi não decepcionar por acreditar em mim.

Aos amigos que me apoiaram, incentivaram e ajudaram, em particular minha irmã Viviane, pela ajuda com as referências em imunologia.

À professora Luciana, por aceitar me orientar nesta jornada, por toda a paciência e atenção dedicadas mim nesses três anos. Ao professor Heder, por ter aceitado o convite de co-orientar este trabalho e por suas contribuições.

Aos professores do Departamento de Ciência da Computação e do Departamento de Mecânica Aplicada e Computacional pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso crescimento pessoal e profissional.

"Existem questões a cuja resposta eu daria um valor infinitamente maior do que às matemáticas, por exemplo questões sobre ética, sobre nosso relacionamento com Deus, sobre nosso destino e nosso futuro; mas a sua resposta encontra-se totalmente além de nós e completamente fora do domínio da ciência" *Carl Frederich Gauss*

Conteúdo

Lista de Figuras **8**

Lista de Tabelas **10**

Lista de Abreviações **13**

1 Introdução **14**

2 Fundamentação Teórica **18**
2.1 Ispiração Natural 18
2.1.1 Sistema Imunológico Natural 18
2.2 Abordagem Computacional 21
2.2.1 Sistema Imunológico Artificial 21
2.2.2 Redes Imunológicas Artificiais (RIAs) 21
2.2.2.1 aiNet 23
2.2.2.2 ARIA 26
2.2.2.3 aiNet X ARIA 31
2.2.2.4 SAINET 33
2.3 Trabalhos relacionados 39
2.3.1 Preservação de densidade em algoritmos imuno-inspirados 40
2.3.1.1 Erro de quantização 40
2.3.1.2 Preservação da densidade 41
2.3.2 Aplicação da ARIA para bases de dados com alta dimensão 42
2.3.3 SAINET com raio adaptativo 43

3 Propostas para a SAINET com raio adaptativo **44**
3.1 Determinar o raio mínimo r 45
3.2 Estimativa da densidade de um anticorpo 45

4 Estudo de caso **47**
4.1 Bases de Dados 47
4.2 Métricas utilizadas na avaliação 51
4.3 Resultados alcançados pelo classificador proposto 53
4.4 Comparações da acurácia 59
4.4.1 Comparação com a SAINET com raio adaptativo 59
4.4.2 Comparação com Naive Bayes 60

5 Conclusões e Trabalhos Futuros **61**

Bibliografia **64**

Lista de Figuras

2.1 Especificidade de anticorpos e agentes patogênicos. À esquerda, temosum exemplo do vírus do Sarampo e os anticorpos produzidos especificamente para se ligarem a ele. À direita, temos os mesmos anticorpos e um tipo do vírus Influenza (vírus da Gripe). Os anticorpos produzidos para o vírus do Sarampo são específicos para ele e não são compatíveis com o vírus da Gripe. Figura extraída de (PARHAM, 2009) 19

2.2 Fases da resposta imunológica adquirida. As três primeiras fases consistem respectivamente em reconhecer o antígeno, ativar os linfócitos e eliminar os antígenos. A resposta sofre redução à medida que parte dos linfócitos estimulados morrem (apoptose). Os linfócitos sobreviventes são matidos para formar a memória imunológica. A duração de cada fase é variável. (ABBAS A. K., 2009) 20

2.3 Identificadores moleculares e teoria da rede. (a) A porção de um antígeno que é reconhecida por um anticorpo é chamada epítopo. Enquanto os anticorpos são monoespecíficos, os antígenos podem apresentar vários epítopos distintos. (b) Molécula de anticorpo destacando paratopo e idiotopo. (c) Respostas positiva e negativa resultantes da interação de um paratopo com um idiotopo ou um epítopo. (KNIDEL, 2006) 22

2.4 Ilustração da aiNet. (a) Conjunto de dados com três *clusters* densos. (b) Rede de anticorpos rotulados e conexões ponderads. As linhas tracejadas indicam conexões a serem removidas para gerar os subgrafos que identificam os *clusters* 24

2.5 Ilustração ARIA. (a) Posicionamento dos anticorpos após a convergência da rede com seus respectivos raios de supressão (raios dos anticorpos). (b) AGM construída sobre os anticorpos. (BEZERRA et al., 2005) 27

2.6 Maturação de afinidade e expansão clonal 30

2.7 Em azul, o raio de vizinhnça (E). Em rosa, o raio dos anticorpos (R). O raio E é utilizado no cálculo da densidade do anticorpo, delimitando aregião de vizinhnça que será considerada. O raio R delimita área de reconhecimento do anticorpo e funciona como limiar para supressão de rede. 31

2.8 Exemplo hipotético da supressão de rede. À esquerda, a rede antes da supressão. O anticorpo B é reconhecido pelo anticorpo A. A prioridade de sobrevivência é dada ao anticorpo com menor raio R, neste caso, o anticorpo B. À direita, a rede após a supressão de rede, que eliminou o anticorpo A. . 32

2.9 Exemplo ilustrativo do citério de partição da AGM. Para este exemplo, quando a aresta CD for avaliada, a menor aresta de sua vizinhança imediata é DF. Para n=2, se o peso de CD for pelo menos duas vezes o peso de DF, CD deverá ser removida. Assim, dois *clusters* serão identificados: (A,B,C) e (D,E,F). (BEZERRA et al., 2005) 32

2.10 População inicial da SAINET para um conjunto de dados com duas classes. As '+' representam os antígenos. Em amarelo, antígenos da classe 1, em azul antígenos da classe 2. O anticorpos são representados por 'o'. Em roxo, o anticopro que representa a classe 1, em vermelho, o anticorpo que representa a classe 2. Neste momento, os dois anticorpos possuem o mesmo raio R definido por um valor aleatório. 37
2.11 Exemplo de comparação do posicionamento dos anticorpos da rede antes e depois do ajuste de pesos. As cruzes representam os antígenos, em amarelo antígenos da classe A e em azul da classe B. Os cículos representam os anticorpos da rede com seus raios, em roxo anticorpos da classe A e em vermelho da classe B. Para este exemplo foi adotado o valor de α =0.5. Observa-se que após o ajuste dos pesos, os anticorpos afastaram-se uns dos outros. Quanto maior o valor de α maior o afastamento. 39

4.1 Base Mk_blobs gerada artificialmente pela ferramenta make_blobs. Cada conjunto corresponde a uma clase. 51

Lista de Tabelas

2.1 Inspiração biológica e aplicações de Sistemas Imunológicos Artificiais . . . 21
2.2 Descrição dos símbolos no algoritmo ARIA 28
2.3 Comparação aiNet X ARIA. 34
2.4 Descrição dos símbolos no algoritmo SAINET. 37

4.1 Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e cálculo original da densidade. Valores obtidos para o raio mínimo r, limiar de reconhecimento de padrões não-próprios (LRPNP), e valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Bases de dados normalizadas através da L2-Norm. 53
4.2 Resultados complementares para as bases de dados com duas classes da aplicação da SAINET com raio adaptativo, com raio mínimo calculado e cálculo original da densidade. Valores médios obtidos para Especificidade (E), sensibilidade (S) e eficiência. 54
4.3 Base Iris: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Parâmetros adotados: r=0,06, LRPNP=0,07, =0,1. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm. 55
4.4 Base Índios Pima: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,28, LRPNP=0,34, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm. 55
4.5 Base Vinhos: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Parâmetros adotados: r=0,014, LRPNP=0,017, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm. 55
4.6 Base BC_Wisconsin: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,37, LRPNP=0,45, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm. 56

4.7 Base Mk_Blobs: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,14, LRPNP=0,18, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm. 56
4.8 Taxas de compressão médias obtidas para as bases de dados em relação ao conjunto de treinamento utilizado adotando o cálculo original da densidade no algoritmo da ARIA. Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 57
4.9 Taxas de compressão médias obtidas para a base da Iris em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 57
4.10 Taxas de compressão médias obtidas para a base dos Índios Pima em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 58
4.11 Taxas de compressão médias obtidas para a base de Vinhos em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 58
4.12 Taxas de compressão médias obtidas para a base BC_Wisconsin relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 58
4.13 Taxas de compressão médias obtidas para a base Mk_Blobs em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média. 58
4.14 Comparação entre os resultados obtidos para as bases da Iris e de Vinhos com os resultados obtidos por (ALMEIDA, 2017) para as mesmas bases de dados. Valores comparados: valores médios da taxa de classificação correta (TCC) e do número de anticorpos (NA). 59

4.15 Valores de r e LRPNP adotados neste trabalho em comparação com os utilizados por (ALMEIDA, 2017), onde os dados não foram normalizados. Comparados com os valores calculados para tais parâmetros considerando as bases sem normalização e com normalização L2-Norm. 59
4.16 Resultados da classifcação usando Gaussian Naive Bayes. 60

Lista de Abreviações

UFJF	Universidade Federal de Juiz de Fora
MAC	Departamento de Mecânica Aplicada e Computacional
DCC	Departamento de Ciência da Computação
IA	Inteligência Artificial
IC	Inteligência Computacional
CE	Computação Evolucionista
SI	Sistema Imunológico
SIA	Sistema Imunológico Artificial
RIA	Rede Imunológica Artificial
aiNet	Artificial Immune Network
ARIA	Adaptative Radius Immune Algorithm
SAINET	Supervised Artificial Immune Network
AGM	Árvore Geradora Mínima
KNN	K-Nearest Neighbors
TCC	Taxa de Classificação Correta
NA	Número de anticorpos
NB	Naive Bayes

1 Introdução

Adaptar-se é o que garante a sobrevivência dos indivíduos. A capacidade de observar, compreender e reproduzir comportamentos que funcionam bem em um determinado ambiente ajuda um indivíduo a se manter por mais tempo no meio em que se encontra. Dos vírus aos mamíferos, organismos têm se adaptado ao longo dos anos às características dos ambientes em que estão inseridos de modo a perpetuar sua espécie passando adiante suas características. Para alguns animais, parte do processo de adaptação passa pela observação e pelo aprendizado. Observar as características do ambiente como clima, vegetação e relevo, e o comportamento dos outros seres vivos como hábitos, habilidades e limitações de suas presas e de seus predadores a fim de agir de forma a se beneficiar e se defender, faz parte da luta pela sobrevivência. Esse processo de observação permite aos indivíduos reconhecer padrões e agir de acordo para se adaptar. Os indivíduos mais adaptados tendem a sobreviver tempo suficiente para se reproduzir e passar adiante suas características, enquanto os menos adaptados são eliminados antes, segundo a Teoria da Seleção Natural proposta pelo naturalista britânico Charles Darwin (DARWIN, 1859) com a finalidade de explicar a Teoria da Evolução das Espécies também proposta pelo mesmo. Reconhecer padrões é o que permite identificar aquilo que é benéfico e o que é prejudicial, como por exemplo diferenciar o alimento do veneno, a presa do predador, o amigo da ameça. Permite também compreender processos e aprimorá-los, tais como as formas de comunicação e relacionamento entre os indivíduos. As técnicas utilizadas para reconhecer padrões variam tanto quanto os próprios padrões e as necessidades atreladas aos mesmos. Cada indivíduo desenvolve sua própria estratégia para alcançar seu objetivo da melhor forma possível, seja este objetivo se alimentar, acasalar ou apenas sobreviver mais um dia. A maneira particular de cada um de fazê-lo inspira estudos para resolução de problemas em diversas áreas da ciência como a Inteligência Artificial (IA), área da Ciência da Computação que se destina ao estudo da síntese e análise dos agentes computacionais que agem de forma inteligente (POOLE; MACKWORTH, 2010).

Segundo (POOLE; MACKWORTH, 2010), um agente é algo ou alguém que age

no ambiente onde está inserido. Agentes inteligentes por sua vez, são agentes que agem de forma inteligente, ou seja, considerando possibilidades e tomando decisões baseados em algum tipo de lógica que faça sentido naquele contexto. Porém, estudar agentes inteligentes por si só por muitas vezes pode não ser suficiente, havendo a necessidade criar novos agentes inteligentes em alguns cenários a fim de melhorar os resultados obtidos. Ainda de acordo com (POOLE; MACKWORTH, 2010), estudar e desenvolver agentes computacionais inteligentes é a objetivo da Inteligência Computacional (IC). A IC se inspira na natureza para elaborar técnicas a fim de desenvolver sistemas inteligentes que imitem aspectos úteis do comportamento natural dos indivíduos como aprendizado, percepção, raciocínio, evolução e adaptação. A área da IC que estuda o comportamento evolutivo e a adaptação dos indivíduos na natureza com interesse computacional é conhecida como Computação Evolutiva (CE) (também chamada de Computação Evolucionista ou Computação Evolucionária), e congrega diversas iniciativas de pesquisa que visam simular aspectos específicos do processo evolutivo (ZUBEN, 2011) A CE se propõe a apresentar soluções aproximadas para problemas de otimização resolvendo-os a partir da descrição matemática da solução objetivo, eliminado a necessidade de especificar os passos tomados até alcançar o resultado (ZUBEN, 2011).

Os algoritmos evolutivos se baseiam na Teoria da Evolução das Espécies e no princípio da seleção natural propostos por Darwin. O problema original é modelado de forma a ser representado por uma população inicial de indivíduos que é evoluída ao longo de um determinado número de gerações objetivando alcançar uma população com as melhores características possíveis a fim de solucionar o problema. A etapa de evolução promove a reprodução do indivíduos, pode ou não impor variáveis aleatórias para diversificar a população, promove competição e seleciona indivíduos da população para compor a próxima geração de indivíduos (ZUBEN, 2011).

Abordagens evolutivas podem ser utilizadas para resolver problemas computacionais diversos, como problemas de otimização, decisão, busca, etc, de diversas áreas de conhecimento. Nos últimos anos, o constante avanço da tecnologia tem gerado volumes de dados cada vez maiores para serem armazenados e analisados, dificultando a extração de informação e conhecimento das bases de dados e demandando novas técnicas para re-

alizar tais processos. O processo para extrair informação dos dados pode ser dividido em seis fases: seleção, pré-processamento, transformação, mineração de dados, avaliação e conhecimento. Este processo é conhecido como KDD (Knowledge Discover in Databases) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 2010) e não é o único processo existente para esta atividade.Parte do processo, a Mineração de Dados pode ser definida como uma área multidisciplinar que utiliza várias técnicas para extrair informações de um conjunto de dados. Minerar os dados consiste em aplicar tarefas de mineração para extrair o conhecimento de forma automática, isto é, automatizando o processo de análise que é feito pelo ser humano a fim de tornar o processo aplicável a volumes de dados maiores sem perda de eficácia. As tarefas mais comuns são agrupamento, associação, classificação, descrição, estimação e predição (CAMILO; SILVA, 2009). Na execução das tarefas podem ser empregados algoritmos que usam diversas abordagens computacionais, que podem inclusive ser combinadas para melhorar os resultados. De um modo especial, a tarefa de classificação consiste em dado um conjunto de dados identificar corretamente a qual classe cada amostra do conjunto pertence, baseando-se nos atributos da amostra. Nesta tarefa, o modelo classificador analisa o conjunto de dados considerando a classe indicada para cada amostra, uma informação conhecida previamente, para encontrar uma relação entre os atributos das amostras e a classe a que pertencem. Dessa forma, o modelo é capaz de classificar novas amostras apresentadas a ele apontando a classe à qual deveriam pertencer.

A motivação para este trabalho vem do gosto pessoal pela IA e pela área de mineração de dados, além do interesse despertado pelo estudo das abordagens evolucionistas em um dado momento da graduação, em particular pelos Sistemas Imunológicos Artificiais (SIAs). A curiosidade de conhecer melhor os SIAs e a dificuldade em encontrar trabalhos relacionados ao tema que trouxessem uma apresentação do mesmo mais compreensível por iniciantes no assunto, motivaram a escrita deste texto sob a proposta de apresentar o trabalho desenvolvido possibilitando que mesmo aqueles que estejam tendo seu primeiro contato com o tema possam compreender os conceitos abordados e formar uma base para iniciar seus próprios estudos. Estas mesmas razões, motivaram o trabalho que foi desenvolvido anteriormente em (ALMEIDA, 2017), onde uma alteração foi

proposta para o algoritmo de uma Rede Imunológica Artificial para classificação que consistia em substituir o algoritmo usado para construção da população de anticorpos de memória (algoritmo da aiNet) pelo algoritmo da ARIA, e foi chamada de SAINET com raio adaptativo. No presente trabalho, alterações sugeridas para o algoritmo da ARIA a fim de melhorar sua qualidade são implementadas com a intenção de avaliar se, estas modificações que sugerem melhoria na qualidade do posicionamento dos anticorpos, impactam a classificação feita pela SAINET com raio adaptativo. Além disto, com o objetivo de reduzir a dependência do algoritmo da SAINET em relação a parâmetros que são definidos pelo usuário, é proposta uma forma de calcular o raio mínimo dos anticorpos e é sugerida uma forma de relacionar o limiar de reconhecimento de padrões não-próprios com raio mínimo calculado.

2 Fundamentação Teórica

Neste capítulo são apresentados de forma breve alguns conceitos e definições necessários para melhor compreensão do tema abordado. São trazidas noções gerais sobre imunologia básica, um resumo dos SIAs e alguns trabalhos relacionados que inspiraram o que é apresentado por este.

2.1 Ispiração Natural

2.1.1 Sistema Imunológico Natural

Para que um organismo funcione de maneira adequada, são necessárias estruturas ou elementos que garantam que as necessidades básicas do mesmo sejam supridas. Além da capacidade de se manter funcionando, o organismo necessita de mecanismos para se defender de ameaças à sua integridade, sejam elas advindas do meio exterior ou do próprio organismo. Historicamente, a palavra imunidade se refere a proteção. No caso dos organismos, a proteção contra agentes que ofereçam risco ao seu funcionamento (agentes patogênicos), sejam estes de origem interna ou externa, é feita por um conjunto de células e moléculas que agem de forma coletiva e coordenada, compondo o Sistema Imunológico (SI). A ação dos componentes do SI é chamada de resposta imunológica (ABBAS A. K., 2009).

A imunidade pode ser dividida em imunidade natural e imunidade adquirida ou adaptativa. A imunidade natural é uma linha inicial de defesa contra microrganismos, composta por mecanismos celulares e bioquímicos que existem antes da exposição do organismo a qualquer infecção e estão programadas para responder rapidamente às ameaças (ABBAS A. K., 2009). São mecanismos de defesa com os quias o indivíduo já nasce, *e.g.* a pele, substâncias antibacterianas presentes na pele, secreções, células especiais como fagocitárias, e algumas proteínas do sangue. A imunidade adaptativa por sua vez, é estimulada pela exposição a agentes patogênicos. Ela se desenvolve em resposta à infecções

e se adapta a elas. A imunidade adaptativa apresenta características como a capacidade distinguir entre o que é próprio do organismo e o que não é, além de recordar de encontros anteriores com um mesmo agente patogênico e reagir mais rapidamente a este em encontros posteriores, evitando assim danos porfundos ao organismo (PARHAM, 2009). Os prinicpais componentes da imunidade adaptativa são células brancas do sangue chamadas linfócitos e seus produtos como os anticorpos. Os linfócitos possuem na superfície da célula receptores para reconhecer agentes patogênicos, o que os tornam mecanismos de defesa mais específicos do que os mecanismos da imunidade inata (PARHAM, 2009). Dessa forma, o anticorpos são capazes de reconhecer a estrutura molecular dos agentes patogênicos e produzir uma resposta imunológica específica para combater este agente. Esta relação é ilustrada na Figura 2.1..

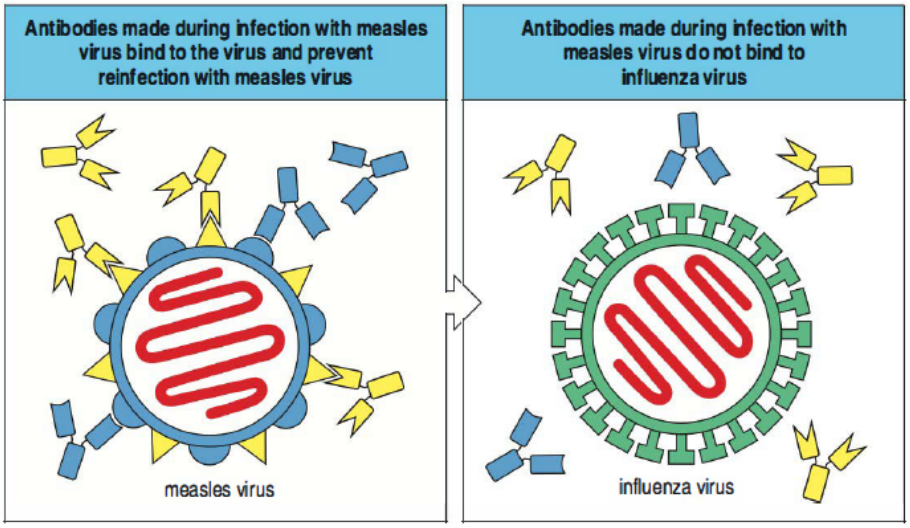

Figura 2.1: Especificidade de anticorpos e agentes patogênicos. À esquerda, temosum exemplo do vírus do Sarampo e os anticorpos produzidos específicamente para se ligarem a ele. À direita, temos os mesmos anticorpos e um tipo do vírus Influenza (vírus da Gripe). Os anticorpos produzidos para o vírus do Sarampo são específicos para ele e não são compatíveis com o vírus da Gripe. Figura extraída de (PARHAM, 2009)

As substâncias estranhas que induzem a resposta imunológica são chamadas antígenos. Durante a infeccção, os antígenos são capturados por células apresentadoras de antígenos, também chamadas APCs, e apresentados a linfócitos específicos. Antígenos que se encontram no interior de células do organismo são apresentados aos linfócitos T que têm a função de reconhecê-los e destruí-los ou destruir células infectadas por eles. Já os antígenos que estão fora das células são apresentados a uma outra classe de linfócitos, os linfócitos B, que são responsáveis por produzir anticorpos para combater tais antígenos.

(ABBAS A. K., 2009) define os anticorpos como proteínas do plasma sanguíneo que se ligam aos antígenos para destruí-los. Anticorpos e antígenos podem ser vistos como figuras que se encaixam, sendo os anticorpos produzidos sob medida, de modo que suas características físico-químicas possibilitem interações com os antígenos.

Ao longo da resposta imunológica, linfócitos receptores são selecionados para atuarem reconhecendo os antígenos (seleção clonal), se diferenciarem uns dos outros e se proliferarem, produzindo uma grande quantidade e variedade de células específicas para aquele antígeno (expansão clonal). Neste processo, anticorpos são selecionados de modo que aqueles que tenham mais chance de combater de maneira eficaz o antígeno sobrevivam, e os linfócitos são aprimorados de modo a produzir anticorpos melhores. Isso faz com que a concentração de linfócitos e anticorpos no sangue aumente por um período, para combater um determinado tipo de antígeno, e caia logo em seguida, quando o antígeno estiver neutralizado. A Figura 2.2 mostra as fases do resposta imunológica.

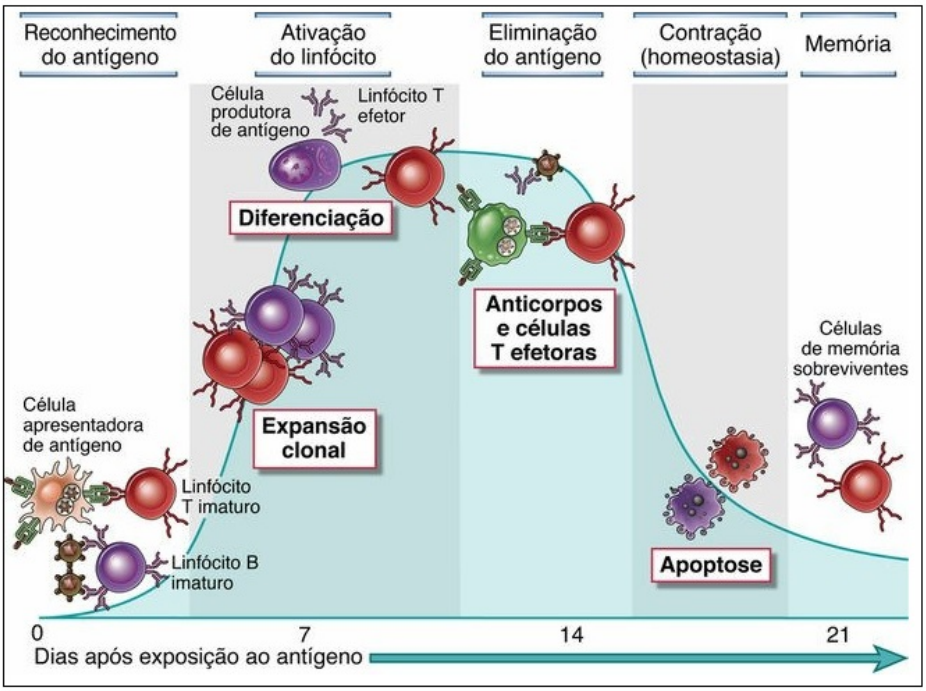

Figura 2.2: Fases da resposta imunológica adquirida. As três primeiras fases consistem respectivamente em reconhecer o antígeno, ativar os linfócitos e eliminar os antígenos. A resposta sofre redução à medida que parte dos linfócitos estimulados morrem (apoptose). Os linfócitos sobreviventes são matidos para formar a memória imunológica. A duração de cada fase é variável. (ABBAS A. K., 2009)

Alguns linfócitos selecionados são mantidos pelo organismo e passam a compor

a chamada mémoria imunológica que permite que, ao ser exposto novamente ao mesmo antígeno no futuro, o SI aja de forma rápida para neutralizá-lo sem passar por todas as etapas da resposta imunológica evitando uma nova infeccção pelo mesmo antígeno. (PARHAM, 2009) Quando isto acontece, os linfócitos responsáveis pela produção destes anticorpos entram em atividade aumentando rapidamente a concentração destes anticorpos neutralizando o antígeno sem grandes dificuldades (LOPES, 2010).

2.2 Abordagem Computacional

2.2.1 Sistema Imunológico Artificial

Inspirados no Sistema Imunológico natural, os Sistemas Imunológicos Artificiais (SIAs) se baseiam em propriedades específicas do sistema natural para resolver problemas computacionais diversos. A Tabela 2.1, extraída de (LOPES, 2010) apresenta uma relação entre a inspiração biológica e os problemas e aplicações típicos abordados. Mais detalhes sobre os tipos de SIAs e suas características podem ser encontrados em (FIGUEREDO; BERNARDINO; BARBOSA, 2013) e em (CASTRO; ZUBEN, 1999).

Aspecto Imunológico	Problema Computacional	Aplicações Tipicas
Reconhecimento de Próprio e Não-Próprio do organismo	Detecção de mudanças ou anomalias	Segurança de computador Detecção de falha
Teoria de Rede Imunológica e Memória Imunológica	Aprendizagem (supervisionada ou não)	Classificação Clusterização Análise de Dados Mineração de Dados
Seleção Clonal	Busca, otimização	Otimização de Funções
Mobilidade e Distribuição	Processamento Distribuído	Arquiteturas de agentes Controle robótico descentralizado
Imunidade Inata	Teoria de Perigo	Segurança de redes

Tabela 2.1: Inspiração biológica e aplicações de Sistemas Imunológicos Artificiais

2.2.2 Redes Imunológicas Artificiais (RIAs)

A Teoria de Rede Imunológica (JERNE, 1974) propõe novas ideias sobre a produção de anticorpos e discute sobre seleção do repertório pré-imune, diferenciação entre próprio e

não-próprio, tolerância ao próprio, evolução e memória do SI. A Teoria de Rede propõe que o sistema está ativo o tempo todo e não precisa de um estímulo externo para começar a agir além de ser capaz de reconhecer também as células e moléculas do próprio organismo. (KNIDEL, 2006).

Em (KNIDEL, 2006) encontramos a definição formal de que o SI, de acordo com a Teoria de Rede, é um conjunto de paratopos e idiotopos que se reconhecem e são reconhecidos entre si. Paratopos e idiotopos são elementos contidos na superfície dos anticorpos, sendo idiotopo um receptor e paratopo um marcador que se liga à superfície de um antígeno em regiões chamadas epitopos. Para a teoria de rede, não só as moléculas são relevantes, mas também as suas interações. Quando uma molécula é reconhecida e identificada como nociva, ocorre o que se chama resposta positiva, resultando em uma proliferação de células, na ativação e secreção de anticorpos para atacar a molécula. Caso contrário, ocorre que se chama resposta negativa, sendo tal molécula tolerada. A Figura 2.1 ilustra esta interação.

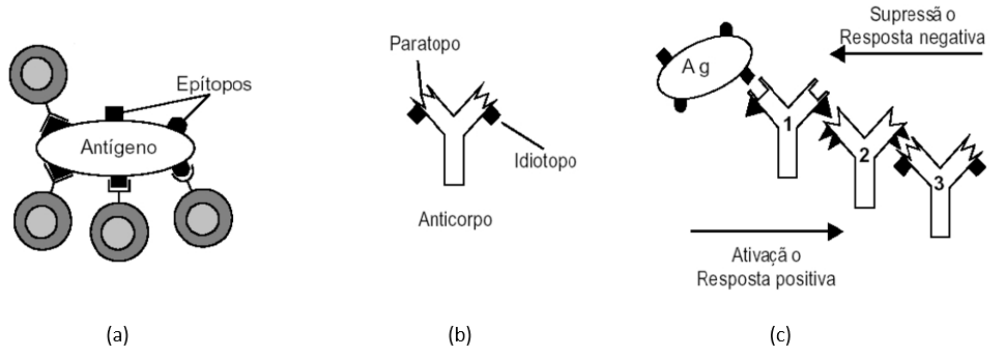

Figura 2.3: Identificadores moleculares e teoria da rede. (a) A porção de um antígeno que é reconhecida por um anticorpo é chamada epítopo. Enquanto os anticorpos são monoespecíficos, os antígenos podem apresentar vários epítopos distintos. (b) Molécula de anticorpo destacando paratopo e idiotopo. (c) Respostas positiva e negativa resultantes da interação de um paratopo com um idiotopo ou um epítopo. (KNIDEL, 2006)

Redes Imunológicas Artificiais (RIAs) são SIAs inspirados fundamentalmente na teoria de rede. As principais características de uma rede imunológica são:

- Estrutura: responsável pela descrição dos padrões de conexão e interação dos componentes da rede
- Dinâmica: variação da concentração e da estrutura dos anticorpos ao longo do tempo

- Metadinâmica: produção contínua de novos anticorpos e morte de anticorpos pouco estimulados. Introduz diversidade e garante eficiência ao SI.

A seguir, são introduzidas três Redes Imunológicas Artificiais, sendo as duas primeiras pra clusterização e a terceira para classificação.

2.2.2.1 aiNet

O primeiro algoritmo para uma RIA foi proposto por (CASTRO; ZUBEN, 2002). Conhecido pelo nome de aiNet (*Artificial Immune Network*), tinha por objetivo inicial resolver problemas de clusterização, e era capaz de identificar automaticamente um número arbitrário de amostras. A aiNet faz uso além da Teoria de Rede, dos princípios de Seleção Clonal e de maturidade de afinidade. A primeira versão do algoritmo ainda introduziu novos aspectos no estudo como as interações da rede entre as soluções e a adaptação dinâmica do conjunto de soluções candidatas.

A partir de 2002, outras versões da aiNet foram desenvolvidas para aplicações em problemas de áreas diversas da computação como classificação, otimização, problemas combinatórios, bioinformática, entre outras. (FRANçA et al., 2010).

A aiNet pode ser formalmente definida como um grafo ponderado, que pode ou não ser totalmente interconectado, composto por um conjunto de nós denominados anticorpos e conjuntos de pares de nós chamados conexões. As conexões possuem um valor característico associado, chamado de peso da conexão ou simplesmente peso (CASTRO; ZUBEN, 2002). A Figura 2.2a mostra o exemplo de um conjunto de dados com 3 *cluster* densos e a Figura 2.2b mostra a arquitetura de uma rede hipotética gerada pelo algoritmo da aiNet para este conjunto, ambas extraídas de (CASTRO; ZUBEN, 2002).

O pseudocódigo é apresentado no Algoritmo 1 e as etapas do algoritmo podem ser resumidas da seguinte forma:

- **Inicialização:**(linha 2) criação de uma população aleatória de anticorpos.

- **Apresentação dos antígenos:** (linhas 5 a 13)

  Seleção clonal e expansão: determinar a afinidade para cada anticorpo da rede, e selecionar um número de anticorpos com alta afinidade para serem clonados de

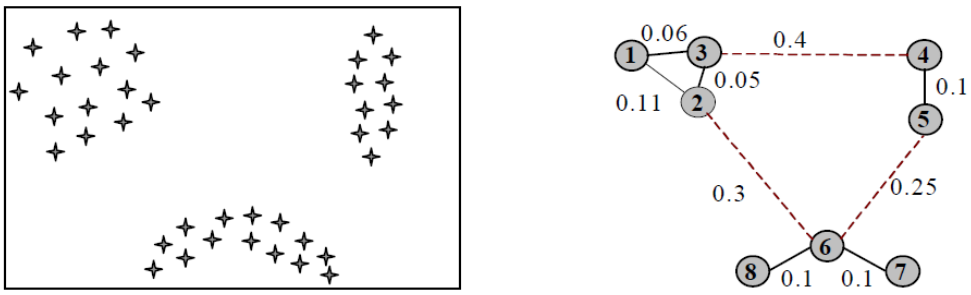

Figura 2.4: Ilustração da aiNet. (a) Conjunto de dados com três *clusters* densos. (b) Rede de anticorpos rotulados e conexões ponderads. As linhas tracejadas indicam conexões a serem removidas para gerar os subgrafos que identificam os *clusters*

forma proporcional à sua afinidade;

Maturação e afinidade: aplicar a cada um dos clones uma mutação inversamente proporcional à sua afinidade;

Morte dos elementos não estimulados: eliminar da memória os clones cuja afinidade com o antígeno seja menor que um limiar estabelecido;

Interações clonais: (linhas 14 e 15) determinar a afinidade de todos os elementos da rede

Supressão clonal: (linha 16) eliminar os clones cuja afinidade entre si seja menor que um limiar estabelecido;

- **Interações da rede:** (linha 17) determinar a similaridade entre cada par de células da rede.

- **Supressão da rede:** (linha 17) eliminar da rede todos os anticorpos cuja afinidade em relação a um anticorpo específico seja maior que um limiar pré-definido. Este anticorpo específico deve ter afinidade a antígenos do que os anticorpos a serem eliminados.

  Construção da rede: (linha 18) incorporar os elementos restantes do conjunto de clones de memória aos anticorpos da rede.

- **Introdução de novos elementos:** (linha 20) introduzir na rede um número fixo de anticorpos gerados aleatoriamente.

- **Ciclo:** repetir os passos 2 a 5 até atingir um número máximo de iterações pré-

estabelecido.

Algoritmo 1: Algoritmo aiNet

Entrada: **N** : tamanho da população inicial de anticorpos
n: número de células a serem clonadas
m: porcentagem de células a serem mantidas no conjunto de memória
σ_d: limiar de supressão anticorpo-antígeno
σ_s: limiar de supressão anticorpo-anticorpo
d: porcentagem de anticorpos gerados aleatoriamente a serem incluidos na população
Saída: população de anticorpos da rede

```
1  início
2      construir população inicial com N anticorpos abs gerados aleatoriamente
3      while não atingir o critério de parada do
4          for cada antígeno ag do
5              determinar a sua afinidade para cada ab da rede;
6              selecionar os n abs com afinidade mais alta;
7              gerarNc clones a partir dos n abs selecionados. Quanto maior a
               afinidade, maior será Nc:
8              Nc = sum_{i=1}^{N} round(N - D_{ab_i,ag_j}.N
9              sendo D_{ab_i,ag_j} a distância entre o anticorpo ab_i e o antígeno ag_j
10             aplicar a hipermutação ao conjunto C de clones gerados, com
               variabilidade inversamente proporcional à afinidade do ab pai
               segundo
11             c*_k = c_k + alpha_k(ag - c_k);
12             onde alpha é a taxa de mutação de cada clone c definida por:
13             alpha_k = e^{-rho*f_i/f_max} ;
14             determinar a afinidade entre ag e cada um dos clones;
15             manter apenas m% dos clones com maior afinidade na
               população;
16             eliminar todos os clones cuja afinidade para com ag seja menor
               que sigma_d (apoptosis);
17             determinar a afinidade entre si dos clones que sofreram mutação
               e eliminar aqueles cuja afinidade seja maior que sigma_s (supressão);
18             concatenar os clones e o conjunto de anticorpos de memória ;
19         end for
20         eliminar da rede abs que não reconheçam nenhum ag;
21         introduzir uma porcentagem d de novos anticorpos na população
22     end while
23 fim
```

A primeira versão da aiNet apresentava algumas características que podem ser aprimoradas para aumentar sua eficiência e estabilidade, e possibilitar a aplicação de RIAs a uma gama maior de problemas. Posteriormente, uma nova rede imuno-inspirada chamada ARIA (*Adaptative Radius Immune Algorithm*) foi proposta por (BEZERRA et

al., 2005) apresentando a capacidade de lidar com a densidade variável de dados entre *clusters*, algo que não era possível com o aiNet.

2.2.2.2 ARIA

Grande parte dos algoritmos imuno-inspirados para agrupamento são baseados em compressão de dados para o volume de dados a ser manipulado, e funcionam posicionando protótipos, chamados de anticorpos, em áreas mais representativas do conjunto de dados (posicionamento de protótipos) e podem ser separados em *clusters* com a ajuda de ferramentas de visualização, o que reduz a complexidade do procedimento. Porém, essa estratégia pode ser problemática em algumas situações, como por exemplo, quando os *clusters* identificados estão muito próximos, quando as densidades de dados variam de *cluster* para *cluster*, ou ainda quando as fronteiras dos *clusters* se sobrepõem. Nesse caso, a compressão de dados pode prejudicar a representação de dados fundamentais para a identificação dos *clusters*. Isso acontece porque o posicionamento de protótipos não leva em conta a densidade dos dados, resultando numa diferença entre a distância entre os anticorpos na representação dos dados pelo modelo e a distância entre dados reais que pode afetar drasticamente a qualidade do particionamento (BEZERRA et al., 2005).

No cenário descrito acima, surge a formulação de um algoritmo imuno-inspirado que usa expansão clonal aliada à informação de densidade presente no conjunto de dados afim de produzir uma representação mais correta dos dados reais, chamado de ARIA (*Adaptative Radius Immune Algorithm*) (BEZERRA et al., 2005). Sendo rápido em termos de execução e simples de compreender e implementar, ARIA faz uso de uma supressão adaptativa do raio (distância entre os anticorpos) que é inversamente proporcional à densidade local para cada vizinho do anticorpo, mantendo assim a aglomeração ou esparsidade dos dados. Preservando a densidade após a compressão, o ARIA gera uma representação mais fiel dos dados em comparação com aiNet, por exemplo, garantindo assim mais eficiência na clusterização (BEZERRA et al., 2005). A representação dos dados por anticorpos e seus raios pode ser vista na Figura 2.3a. Nela, é possível ver a rede construída pelo algoritmo para uma base de dados com dois *clusters* com densidades diferentes. A Figura 2.3b mostra a Àrvore Geradora Mínima (AGM) obtida sobre a

população de anticorpos, que deve ser particionada a fim de identificar os *clusters*.

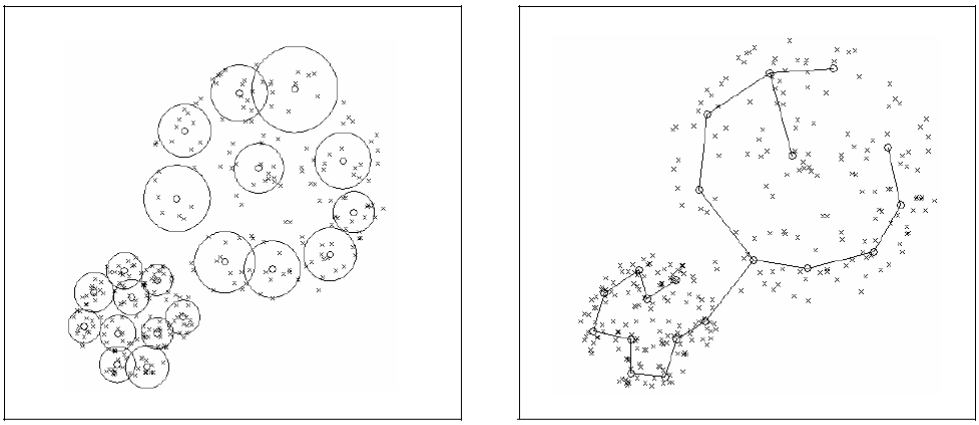

Figura 2.5: Ilustração ARIA. (a) Posicionamento dos anticorpos após a convergência da rede com seus respectivos raios de supressão (raios dos anticorpos). (b) AGM construída sobre os anticorpos. (BEZERRA et al., 2005)

ARIA é um algoritmo adaptativo que pode ser divido em três etapas (BEZERRA et al., 2005):

1. Maturação de Afinidade (mutação): os dados, representados por antígenos, são apresentados aos anticorpos. Os anticorpos sofrem hipermutação (mecanismo de mutação que faz as células do SI se adaptarem a novos invasores) a fim de se adaptarem aos antígenos (interação antígeno-anticorpo);

2. Expansão Clonal: os anticorpos estimulados são selecionados para serem clonados, expandindo assim a rede;

3. Supressão de Rede: a interação entre anticorpos é quantificada e, se um anticorpo reconhece o outro, um deles é removido do *pool* de células (interação anticorpo-anticorpo).

O Algoritmo 2 apresenta o pseudocódigo do ARIA e uma descrição dos símbolos utilizados (BEZERRA et al., 2005) na Tabela 2.2.

A inicialização das variáveis é feita definindo valores para os parâmetros de entrada do algoritmo e definindo a população inicial. Os parâmetros R e E podem ser inicializados com valores aleatórios, uma vez que serão calculados e ajustados pelo algoritmo ao longo das gerações. Para a taxa de mutação μ é sugerido o valor igual a 1, ao

Algoritmo 2: ARIA

Entrada: conjunto de dados, r, gen, μ, *decay*
Saída: população de anticorpos da rede

```
1  início
2      inicializar variáveis;
3      for cada antígeno Ag do
4          selecionar melhor anticorpo Ab;
5          aplicar operador de mutação em Ab segundo mi;
6      end for
7      matar anticorpos não estimulados;
8      clonar anticorpos que reconhecem antígenos a uma distância maior que
       R;
9      calcular a densidade local para cada Ab;
10     calcular calcular o limiar de supressão de Ab fazendo:
       R_Ab = r * (den_max/den)^(1/dim)
11     suprimir anticorpos dando prioridade de sobrevivência para aqueles
       com menor R;
12     Fazer E = media(R);
13     if geração corrente > gen/2 then
14         μ = μ * decay;
15     end if
16 fim
```

Símbolo	Descrição
R	Raio de cada anticorpo (limite de supressão)
r*	Multiplicador do raio. Determina o raio mínimo dos anticorpos
μ	Taxa de Mutação
*decay**	Multiplicador constante usado para diminuir a taxa de mutação
E	Raio que define a vizinhança para estimativa de densidade
gen*	número de iterações
dim	Dimensão dos dados de entrada

Tabela 2.2: Descrição dos símbolos no algoritmo ARIA

longo das gerações este valor decai de acordo com a taxa de decaimento *decay* que pode ser escolhida pelo usuário. A população inicial pode ter qualquer tamanho, desde que tenha pelo menos 1 anticorpo, e as componentes do vetor de atributos de cada anticorpo da população inicial podem ter valores aleatórios, uma vez que o tamanho da população e os atributos dos anticorpos serão ajustados também ao longo do processo evolutivo.

Na maturação de afinidade, cada um dos antígenos é comparado com todos os anticorpos que compõem a população atual, sendo selecionado para sofrer mutação o anticorpo que possui maior afinidade com o antígeno, onde a afinidade é definida como a

menor distância euclidiana (pode-se aplicar qualquer outra métrica ou cálculo de distância compatível com o tipo dos dados do problema em questão). Para cada anticorpo selecionado, é feita uma cópia, antes da mutação, que poderá ou não ser clonada mais a frente. Os anticorpos que sofreram mutação compõem a população intermediária construída e manipulada a cada geração. O operador mutação é definido como segue na Equação 2.1, onde Ab' é o anticorpo resultante da mutação, Ab é o anticorpo selecionado,μ é a taxa de mutação, *rand* é um valor aleatório uniformemente gerado entre 0 e 1, Ag é o antígeno considerado, e *dist*(Ag,Ab) é a Distância Euclidiana entre Ag e Ab.

$$Ab' = Ab + \mu \times rand \times dist(Ag, Ab) \tag{2.1}$$

Os anticorpos que não foram estimulados nesta fase são eliminados, ou seja, não integram a população intermediária (supresão de anticorpos).

Em seguida começa a expansão clonal. Nesta fase, as cópias de cada anticorpo selecionado para mutação são avaliadas. Ressalta-se que as cópias são feitas antes da mutação, sendo portanto cópias idênticas do anticorpo pai. Além disso, para cada anticorpo estimulado só é permitida uma única cópia. Tendo em vista que um mesmo anticorpo pode ser estimulado por antígenos diferentes, a clonagem é feita na direção do primeiro antígeno a estimulá-lo. O operador de clonagem é definido exatmente como o operador de mutação apesentado na equação 2.1.

A clonagem é feita sobre a cópia do anticorpo pai, e acontece se e somente se a distância entre o anticorpo e o antígeno for maior do que o raio de supressão do anticorpo. Caso ocorra a clonagem, o clone é incorporado à população intermediária.

Em resumo, um clone é uma cópia mutada, que pode ou não ocorrer, de um anticorpo estimulado na Maturação de Afinidade. A Figura 2.6 traz uma representação esquemática da maturação se afinidade e da expansão clonal.

Por fim, ocorre a supressão de rede. Esta é a última fase do algoritmo e consiste em eliminar da população intermediária anticorpos que se reconheçam. Para isso, é necessário atualizar o raio R de cada anticorpo. A densidade de cada anticorpo é definida

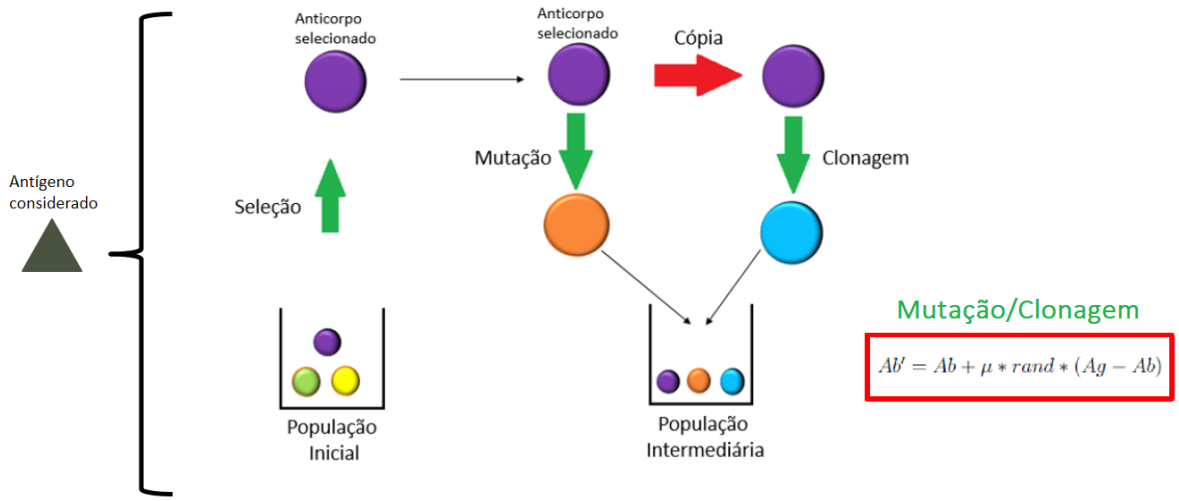

Figura 2.6: Maturação de afinidade e expansão clonal

como a quantidade de antígenos localizados na região delimitada pelo raio de vizinhança E do anticorpo. O raio R do anticorpo é então atualizado em função da sua densidade, de acordo com a Equação 2.2. Na Figura 2.7, mostram-se, para um exemplo hipotético, o raio de vizinhança E e os raios R dos anticorpos calculados pela Equação 2.2. A Figura 2.8 ilustra a supressão de rede para o mesmo exemplo.

$$R_{ab} = r \times (den_{max}/den_{ab})^{(1/dim)} \tag{2.2}$$

onde R_{ab} é o raio do anticorpo ab, r é o raio mínimo (menor valor que R_{ab} pode assumir), den é a densidade do anticorpo ab.

A identificação dos *clusters* ocorre de forma semelhante à da aiNet. Depois de construída a população de anticorpos da rede, uma AGM ponderada é induzida sobre as amostras representados pelos anticorpos, onde os pesos das arestas são as distâncias entre os anticorpos conectados por elas. O particionamento ocorre segundo o seguinte critério: uma aresta e será removida se seu peso for pelo menos n vezes o peso da aresta com menor peso de sua vizinhaça imediata, ou seja, a menor de todas arestas que incidem nos nós conectados por e. Em geral, adota-se n igual a 2. (BEZERRA et al., 2005). A Figura 2.9 apresenta um exemplo ilustrativo do critério de particionamento.

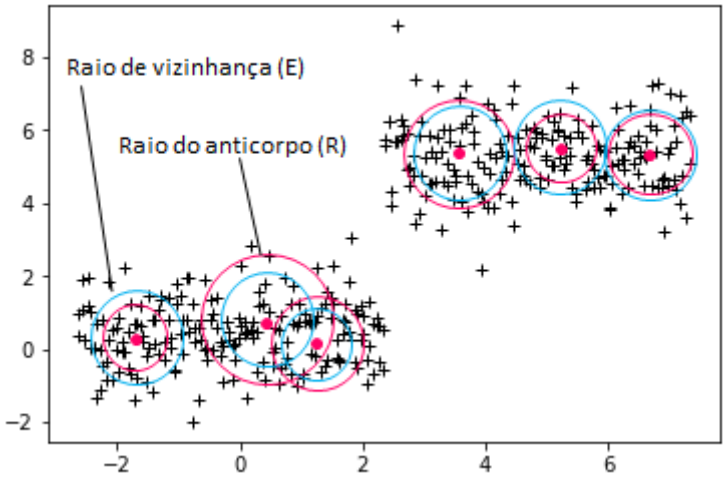

Figura 2.7: Em azul, o raio de vizinhnça (E). Em rosa, o raio dos anticorpos (R). O raio E é utilizado no cálculo da densidade do anticorpo, delimitando aregião de vizinhnça que será considerada. O raio R delimita área de reconhecimento do anticorpo e funciona como limiar para supressão de rede.

2.2.2.3 aiNet X ARIA

Em termos dos parâmetros de cada algoritmo, dos oito parâmetros importantes da aiNet, seis (parâmetros de entrada, além do critério de parada) são definidos pelo usuário, e apenas dois são calculados em função da população (taxa de mutação α e o tamanho Nc do conjunto de clones), o que torna o algoritmo muito sensível aos parâmetros de entrada. Além disso, não está claro em (CASTRO; ZUBEN, 2002) como definir a taxa de mutação α, utilizada no processo de maturação de afinidade dos anticorpos (linhas 10 a 13 do algoritmo).

Já no algoritmo ARIA, dos sete parâmetros indicados na Tabela 2.2, apenas o raio mínimo r e o fator de decaimento da taxa de mutação *decay* (além do critério de parada, que neste caso é número de gerações *gen*) são definidos pelo usuário. Informações sobre como definir o valor de r podem ser encontradas em (AZZOLINI; VIOLATO; ZUBEN, 2010).

Ainda sobre os parâmetros, é interessante atentar para os limiares de supressão adotados por ambos. A aiNet adota um limiar de supressão σ_d para eliminar anticorpos com baixa afinidade para com os antígenos, fase chamada de *apoptosis*, e um outro limiar σ_s para eliminar os anticorpos que tenham alta afinidade com outros anticorpos, ou seja, que reconheçam outros anticorpos, fase conhecida como supressão de rede. A apoptosis

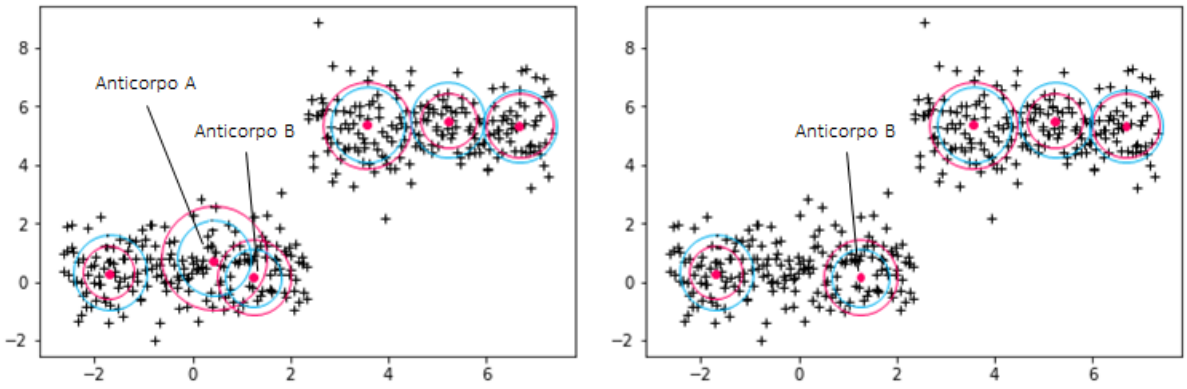

Figura 2.8: Exemplo hipotético da supressão de rede. À esquerda, a rede antes da supressão. O anticorpo B é reconhecido pelo anticorpo A. A prioridade de sobrevivência é dada ao anticorpo com menor raio R, neste caso, o anticorpo B. À direita, a rede após a supressão de rede, que eliminou o anticorpo A.

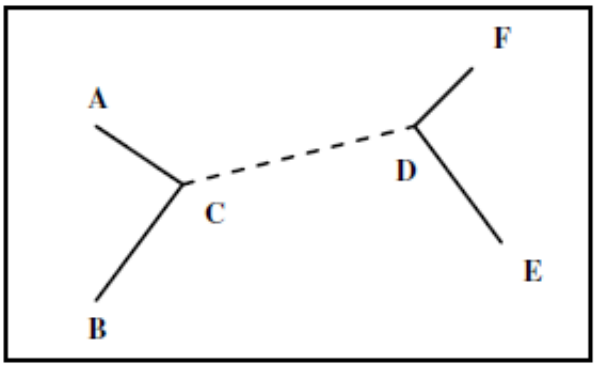

Figura 2.9: Exemplo ilustrativo do citério de partição da AGM. Para este exemplo, quando a aresta CD for avaliada, a menor aresta de sua vizinhança imediata é DF. Para n=2, se o peso de CD for pelo menos duas vezes o peso de DF, CD deverá ser removida. Assim, dois *clusters* serão identificados: (A,B,C) e (D,E,F). (BEZERRA et al., 2005)

no algoritmo da ARIA é feita suprimindo os anticorpos que não foram estimulados por nenhum antígeno durante a maturação de afinidade, não havendo portanto a necessidade de um limiar para decidir se anticorpo é ou não mantido. A supressão de rede adota de forma semelhante um limiar de supressão, que é o próprio raio do anticorpo, sendo este calculado pelo próprio algoritmo em função do parâmetro r e da densidade que é dada em função do raio de vizinhança E que também é atualizado ao longo das iterações.

A mutação se dá de forma semelhante em ambos algoritmos: aplicando-se um operador mutação sobre o anticorpo selecionado levando em conta a afinidade para com o antígeno, uma taxa de mutação e os atributos de ambos. Por outro lado, o processo de clonagem ocorre de forma bem distinta na ARIA em relação à aiNet. Enquanto a clonagem na aiNet ocorre antes da mutação e se resume a gerar cópias idênticas do anticorpo

pai, na ARIA a clonagem consiste em aplicar o operador de mutação novamente à uma cópia feita do anticorpo pai no momento em que este é selecionado para sofrer mutação, caso o antígeno esteja localizado a uma distância do anticorpo maior que seu raio. Em outras palavras, cada clone nada mais é do que uma cópia mutada de seu anticorpo pai (BEZERRA et al., 2005). A ARIA limita o número de clones permitido por anticorpo a, no máximo 1. Desta forma, limita-se a quantidade possível de descendentes de cada anticorpo: uma cópia fiel do mesmo mantida na próxima geração, uma versão mutada e (ou não) uma cópia mutada do anticorpo, caso este seja estimulado. Enquanto na aiNet, o número de clones de cada anticorpo é proporcional à sua afinidade para com o antígeno considerado (linha 8 do Algoritmo 1).

A saída de ambos algoritmos tem o mesmo formato: um conjunto de anticorpos de memória que representam o conjunto de dados (antígenos) de forma reduzida no número elementos, porém sem perda significante de informação.

A Tabela 2.3, apresenta de forma resumida, a comparação entre os algoritmos da aiNet e da ARIA.

2.2.2.4 SAINET

A partir da versão original do aiNet, várias adaptações do algoritmo destinadas a resolver problemas de classificação foram desenvolvidas. Além destas, novas propostas de redes neuro-imunológicas, redes imunológicas híbridas de Redes Neurais Artificiais (RNAs) com intuito de aprimorar os resultados. Alguns exemplos de redes supervisionadas para classificação são apresentados e discutidos por (KNIDEL, 2006). A Supervised Artificial Immune Network (SAINET) é um algoritmo para classificação de dados desenvolvido a partir do aiNet original. O funcionamento da SAINET é constituído de três fases (KNIDEL, 2006):

1. **Aplicação do aiNet ao conjunto original de dados analisado para construir um conjunto de memória que reconheça e represente a organização estrutural dos dados:** o número de anticorpos é controlado por um limiar de supressão relacionado à especificidade dos anticorpos da rede. Por ser uma abordagem supervisionada, há necessidade de rotular os anticorpos responsáveis pela classificação,

	aiNet	ARIA
Parâmetros de entrada (definidos pelo usuário)	σ_d, σ_s, N, m, d, gen	r,*decay*, gen
Parâmetros calculados pelo algoritmo	Nc, α	E, R, μ, dim
Limiares de supressão	σ_d, σ_s	R
Mutação	aplica-se um operador Mutação sobre o anticorpo selecionado levando em conta a afinidade para com o antígeno, uma taxa de mutação e os atributos de ambos	aplica-se um operador Mutação sobre o anticorpo selecionado levando em contaa afinidade para com o antígeno, uma taxa de mutação e os atributos de ambos
Clonagem	Ocorre antes da mutação; gera cópias idênticas ao anticorpo pai em quantidade proporcional à afinidade do anticorpo para com o antígeno considerado	Ocorre depois da mutação; aplicação do operador de mutação em uma cópia idêntica do anticorpo pai e se limita a, no máximo, 1 clone por anticorpo selecionado
Supressão clonal (apoptosis)	Os clones cuja afinidade para com o antígeno considerado seja menor que d são removidos	Os anticorpos que não foram estimulados durante a Maturação de afinidade são eliminados da população intermediária
Supressão de rede	Elimina da rede anticorpos que tenham afinidade para com outros anticorpo menor que σ_s e os anticorpos que não reconhecem nunhum atígeno	Elimina da rede anticorpos que se reconhecem, ou seja, anticorpos localizados a uma distância menor que o raio de algum deles

Tabela 2.3: Comparação aiNet X ARIA.

utiliza-se então um vetor para armazenar o rótulo ao qual cada anticorpo pertence e, quando um anticorpo é clonado na expansão, o rótulo do clone é o mesmo do anticorpo que o gerou.

2. **Verificação da necessidade de duplicação do anticorpo:** feita através de uma heurística que apura se os anticorpos são capazes de reconhecer uma quantidade de padrões não-próprios maior que um limiar estabelecido pelo usuário (LRPNP – Limiar de Reconhecimento de Padrões Não-Próprios). Em outras palavras, avalia a capacidade de reconhecer padrões diferentes da classe à qual o anticorpo pertence. Caso essa condição seja satisfeita, o anticorpo é duplicado, sendo que, os atributos do novo anticorpo são a média aritmética das amostras pertencentes à classe considerada.

3. **Atualização dos pesos:** nesta fase se encontra a aprendizagem supervisionada. Os pesos da rede são atualizados de forma semelhante à adotada na Aprendizagem por Quantização Vetorial (KOHONEN, 2010), segundo a equação 3.3, apresentada na seção 3.2.1,

O Algoritmo 3 apresenta o pseudocódigo da SAINET e uma breve descrição dos parâmetros utilizados no código (KNIDEL, 2006) é mostrada na Tabela 2.3.

A primeira fase caso do algoritmo (linhas 1 a 13) constrói a população de anticorpos da rede. Para isto, aplica-se o algoritmo da aiNet com o objetivo de gerar uma população de anticorpos posicionados de forma a representar a base de dados. Nesta fase, o algoritmo da aiNet não é executado até o final, não sendo executadas as fases refrentes à identificação dos *clusters*. Na segunda fase (linhas 15 a 20) é aplicada a herística que verifica a necessidade de duplicação dos anticorpos gerados na primeira fase. Por fim, na terceira fase (linhas 22 a 25) a posição dos anticorpos é ajustada considerando a informação de classe. É aplicada uma perturbação na solução que afasta os anticorpos uns dos outros de forma proporcional à taxa de aprendizado α

O algoritmo da SAINET é detalhado como segue:

1. **Inicialização das variáveis:**

A população inicial contém um anticorpo um anticorpo para cada classe presente

Algoritmo 3: SAINET.

Entrada: conjunto de dados,gen,Nc,Ts,Tr,α,σ
Saída: AB,R,

início
  inicializar a rede com um anticorpo ab para cada classe presente nos dados de treinamento com a média dos atributos da classe da classe a qual ele for atribuído;
  while *iteração for menor que máximo de gerações gen* **do**
    for *cada antígeno ag* **do**
      apresentar um antígeno escolhido aleatóriamente à rede, sem repetição ;
      calcular a distância euclidiana entre o ag apresentado e todos os abs da rede;
      clonar o anticorpo vencedor (o mais próximo do ag apresentado) produzindo Nc clones ;
      atualizar o vetor de atributos do anticorpo com
      $ab_k = ab_k + \alpha * (ag - ab_k)$
      onde k é o índice do ab vencedor e é a taxa de aprendizagem;
    end for
    atualizar o fator de aprendizagem: $= \sigma * \alpha$;
    suprimir os anticorpos que se auto reconhecem dentro do limiar de supressão Ts
  end while
  for *para cada anticorpo ab* **do**
    for *cada classe diferente da classe de ab* **do**
      if *razão entre o número de padrões que reconhecidos por ab que peretencem a uma classe diferente da sua, pelo número de padrões que reconhecidos por ab que peretencem à mesma classe de ab for maior que Tr* **then**
        duplicar ab e atribuir à cópia a média dos atributos da classe de ab
      end if
    end for
  end for
  while *iteração menor que o número máximo de gerações gen* **do**
    apresentar um antígeno agescolhido aleatoriamente à rede, sem repetição;
    encontrar o anticorpo ab vencedor e atualizar seu vetor de atributos de acordo com a equação 3.3
  end while
fim

no conjunto de treinamento, de modo que todas as classes estejam representadas na população. Os atributos destes anticorpos são definidos como a média dos atributos da classe que representam. A Figura 2.10 mostra a população inicial da rede.

Símbolo	Descrição
gen	número de gerações
Nc	número de clones gerados
Ts	limiar de supressão usado na eliminação de anticorpos que se auto-reconhecem
Tr	limiar de reconhecimento de padrões não-próprios
α	taxa de aprendizagem
σ	fator que controla o decaimento da taxa de apremdizagem
Matriz |AB|	conjunto de anticorpos (protótipos)
Vetor |R|	armazena os rótulos (classes) dos anticorpos da matriz |AB|

Tabela 2.4: Descrição dos símbolos no algoritmo SAINET.

2. **Fase 1:** (linhas 1 a 13)

   A primeira fase consiste na aplicação da aiNet, apresentada na Seção 2.1.3, para construir um conjunto de anticorpos de memória. A execução do algoritmo da aiNet neste caso se limita a gerar a população de anticorpos que será aprimorada nas fases seguintes da SAINET, não sendo necessárias as etapas da aiNet que fazem a identificação dos clusters.

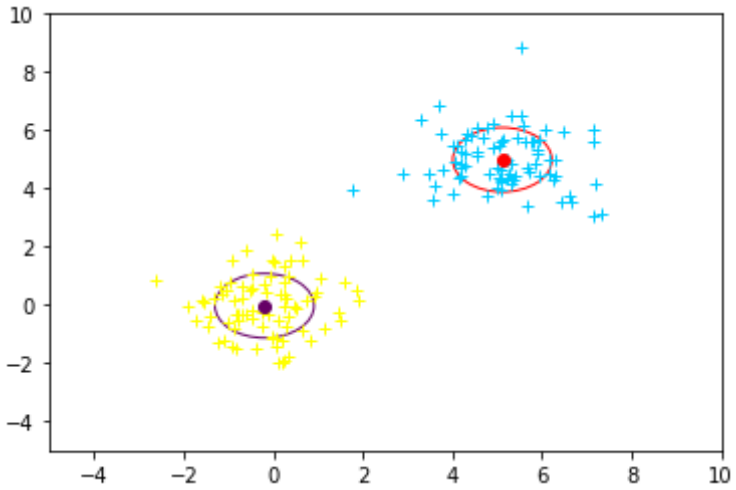

Figura 2.10: População inicial da SAINET para um conjunto de dados com duas classes. As '+' representam os antígenos. Em amarelo, antígenos da classe 1, em azul antígenos da classe 2. O anticorpos são representados por 'o'. Em roxo, o anticopro que representa a classe 1, em vermelho, o anticorpo que representa a classe 2. Neste momento, os dois anticorpos possuem o mesmo raio R definido por um valor aleatório.

3. **Fase 2:** (linhas 15 a 20)

   A população gerada na fase anterior é submetida à uma heurística que verifica a necessidade de duplicação de cada anticorpo com base nos padrões reconhecidos por ele. Cada anticorpo desde o início já tem a informação sobre a classe a qual pertence, uma vez que a população inicial foi construída com um representante de cada classe presente no conjunto de dados de treinamento e os anticorpos filhos mantém o rótulo de classe do pai nas operações de clonagem e mutação. Dessa forma, sendo conhecidas também as classes de cada antígeno, calcula-se a razão entre o número de antígenos reconhecidos pelo anticorpo que pertencem à mesma classe deste, e número de antígenos reconhecidos pelo anticorpo que pertencem à classes diferentes. Caso esta razão exceda o valor definido para o Limiar de Reconhecimento de Padrões Não-Próprios (LRPNP) definido, o anticorpo é então duplicado sendo os atributos do novo anticorpo gerado a média dos atributos da classe a qual o original pertence. Se o anticorpo reconhecer padrões de mais de uma classe além da sua, esta análise deve ser feita de forma independente para cada classe envolvida. Vale reforçar que neste contexto, reconhecer significa estar contido na região do espaço delimitada pelo raio do anticorpo, ou seja, um antígeno *ag* é reconhecido por um anticorpo *ab* se *ag* estiver dentro da região delimitada pelo raio de *ab*.

4. **Fase 3:** (linhas 22 a 25)

   A última fase do algoritmo constitui o aprendizado supervisionado da rede de fato. Os pesos, neste caso, os atributos dos anticorpos são atualizados pela equação 3.3

$$Ab_k(t+1) = \begin{cases} Ab_k(t) + \alpha \times (Ag - Ab_k(t)), \text{ se } Classe(Ab_k) = Classe(Ag) \\ Ab_k(t) - \alpha \times (Ag - Ab_k(t)), \text{ caso contrário} \end{cases} \quad (2.3)$$

   onde Ab_k é o anticorpo em análise, Ag é um antígeno apresentado, e *aplha* é a taxa de mutação.

A atualização dos pesos causa uma espécie de perturbação na rede, movendo anticorpos de lugar. Esta perturbação tende a melhorar o posicionamento dos anticorpos de um modo geral, e corrigir também o posicionamento dos anticorpos gerados na segunda fase caso tenha ocorrido duplicação de algum anticorpo. A intesidade da perturbação depende do valor da taxa de aprendizado α. Quanto maior o valor de α para mais longe os anticorpos serão movidos, o que influencia no resultado do classificador. A Figura 2.11 compara o posicionamento dos anticorpos antes e depois da atualização dos pesos. Observa-se que após o ajuste, os anticorpos afastaram-se uns dos outros diminuindo a intercessão entre eles e aumentando a área coberta pelos anticorpos.

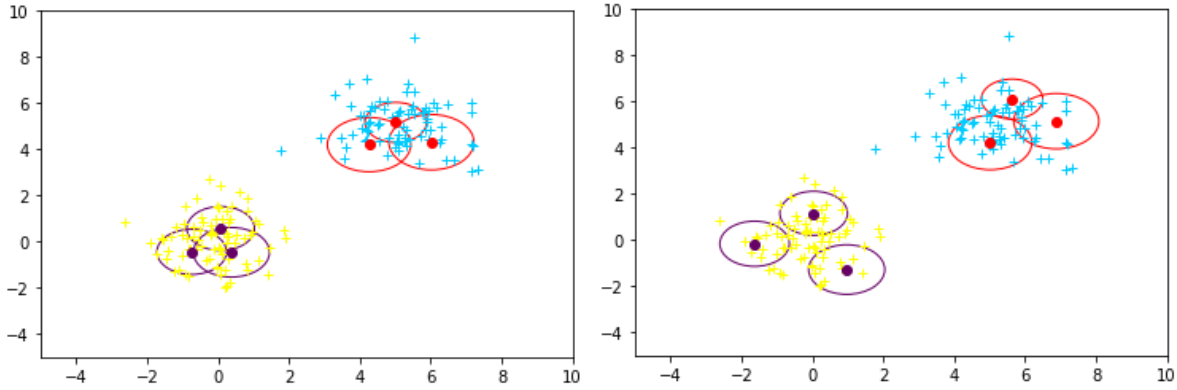

Figura 2.11: Exemplo de comparação do posicionamento dos anticorpos da rede antes e depois do ajuste de pesos. As cruzes representam os antígenos, em amarelo antígenos da classe A e em azul da classe B. Os cículos representam os anticorpos da rede com seus raios, em roxo anticorpos da classe A e em vermelho da classe B. Para este exemplo foi adotado o valor de α =0.5. Observa-se que após o ajuste dos pesos, os anticorpos afastaram-se uns dos outros. Quanto maior o valor de α maior o afastamento.

2.3 Trabalhos relacionados

Esta seção traz alguns trabalhos relacionados que inspiraram o que é apresentado nos capítulos seguintes.

2.3.1 Preservação de densidade em algoritmos imuno-inspirados

Algoritmos de agrupamento que se baseiam em posicionar protótipos em relação aas amostras da base de dados com a finalidade de usá-los para representar os dados precisam minimizar a distância cumulativa entre os protótipos e os dados, chamada erro de qauntização (*minimum quantization error*). Todavia, para algumas aplicações é necessário preservar a densidade dos dados ao posicionar os protótipos, isso significa que os protótipos posicionados devem obedecer o máximo possível a distribuição de densidade dos dados. Entretanto, (AZZOLINI; VIOLATO; ZUBEN, 2010) demonstra que estes dois critérios são possíveis de ser atendidos de forma distinta mas não é possível atendê-los simultâneamente.

Minimizar o erro de quantização se resume a minimizar o somatório das distâncias entre cada ponto da base de dados (antígeno) e seu protótipo (anticorpo) mais próximo. Já a preservação da densidade pode ter diferentes significados dependendo de como a densidade das amostras é estimada e do método utilizado para comparar a densidade estimada dos dados (antígenos) e dos protótipos (anticorpos). O erro de quantização e a preservação da densidade podem ser utilizados como métricas para avaliar a compressão dos dados, e as formulações matemáticas adotadas por (AZZOLINI; VIOLATO; ZUBEN, 2010) para ambos são apresentadas a seguir.

2.3.1.1 Erro de quantização

Dado um espaço finito $\Omega \subset \mathbb{R}^l$, considere um conjunto de dados de entrada N com n vetores $x_i \in \Omega$, e o conjunto de protótipos M com m protótipos $z_i \in \Omega$, onde m ¡ n. Conhecendo M, qualquer vetor $x \in \Omega$ pode ser aproximado pelo protótipo mais próximo em M. Ou seja, M partições Ω em m politopos de Voronoi, cada um correspondendo a uma região representada pelo protótipo. Utilizando uma métrica $d(.,.)$ que retorne uma medida da distorção entre dois vetores, o erro de quantização Q_i entre um vetor $x_i \in N$ e sua versão quantizada $x_i \in M$ é definido por:

$$Q_i = d^r(x_i, q(x_i)) \tag{2.4}$$

onde o expoente r>0 representa a importância dada uma certa distância entre um ponto do conjunto de dados e seu protótipo. Em geral, adota-se r=2 para simplificar o algoritmo.

O erro de quantização médio que é utilizado como métrica é definido como a média aritmética das distâncias de cada ponto em N pelo protótipo mais próximo em M:

$$Q_N = \frac{1}{n}\sum_{i=1}^{n} Q_i = \frac{1}{N}\sum_{i=1}^{n} d^r(x_i, q(x_i)) \tag{2.5}$$

2.3.1.2 Preservação da densidade

A definição usada por (AZZOLINI; VIOLATO; ZUBEN, 2010) de preservação de densidade é a mesma usada por (STIBOR; TIMMIS, 2007) e é usada para avaliar a qualidade da compressão da aiNet. As distribuições de probabilidade dos dados e dos protótipos é estimada por um método não paramétrico considerando que cada ponto dos conjuntos N e M é selecionado como amostra de forma independente de sua distribuição de probabilidade. Para medir a dissimilaridade entre as duas funções estimadas de densidade de probabilidade, é usada estimativa de Monte Carlo da entropia relativa (também chamada divergência de Kullback-Leibler).

Tendo conhecidas as densidades de probabilidades p_N de N e p_M de M, é possível medir a dissimilaridade H$\{N,M\}$ entre os conjuntos pela entropia relativa dada por:

$$H\{N,M\} = \int_{\Omega} \left[\frac{p_N(x)}{p_M(x)}\right] p_N(x) d(x) \tag{2.6}$$

Quanto mais distintos os conjuntos, maior será o valor de H, sendo **H**=0 quando os conjuntos forem idênticos. Conhecendo o número de amostras $x_i \in N$, para i = 1,. . . ,n, o seguinte aproximação de Monte Carlo pode ser usada:

$$\hat{H}\{N,M\} = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{p_N(x)}{p_M(x)} \tag{2.7}$$

Como as funções pN(x) e pM(x) na maioria dos casos não são conhecidas, é possível estimá-las usando aproximações não paramétricas como a feita pelo método do k-ésimo vizinho mais próximo (KNN - K-Nearest Neighbors) por exemplo. O KNN estima

a densidade em torno de um ponto **x** através da equação:

$$p(\mathbf{x}) = \frac{k}{c_l[d_k(x)]^l} \tag{2.8}$$

onde l é a dimensão dos dados de entrada, c_l é o volume da hiperesfera unitária em $\mathbb{R}^l$, e d_k é a distância de x para seu k-ésimo vizinho mais próximo.

2.3.2 Aplicação da ARIA para bases de dados com alta dimensão

Ao lidar com bases de dados com 2 ou 3 dimensões, é possível visualizar graficamente a distribuição dos dados e o posicionamento dos anticorpos feito por algoritmos como da ARIA. Quando em dimensões mais altas, não é possível visualizar a distribuição dos dados nem o comportamento destes algoritmos. Nesses casos, é necessário utilizar de outras ferramentas para avaliar o posicionamento dos anticorpos. Além disso, a compressão de dados não preserva a densidade dos dados quando estes possuem muitos atributos (VIOLATO; AZZOLINI; ZUBEN, 2010), e os cálculos de distância em altas dimensões é complicado uma vez que a distância relativa entre dois amostras tende a zero à medida que o número de dimensões aumenta para algumas distribuições de dados (BEURER, 2010).

Neste contexto, (VIOLATO; AZZOLINI; ZUBEN, 2010) propõe uma modificação no algoritmo da ARIA para preservar a densidade quando aplicado a bases de dados de grande dimensão. A modificação proposta consiste em alterar a forma como a ARIA determina a densidade dos anticorpos para fazer o ajuste do raio de cada anticorpo. O algortimo original calcula a densidade contabilizando os antígenos localizados na região delimitada pelo raio de vizinhaça *E*. a proposta dos autores é que, ao invés de ser calculada, a densidade seja estimada da forma sugerida por (AZZOLINI; VIOLATO; ZUBEN, 2010). Como métrica de avaliação da compressão, é utilizada a entropia relativa assim como sugerido por (AZZOLINI; VIOLATO; ZUBEN, 2010).

A estimativa da densidade pode ser feita com base em métodos não paramétricos como estimador Kernel ou o KNN. A proposta dos autores é estimar a densidade em torno

de um dado anticorpo pela Equação 2.8, e avaliar a qualidade da compressão dos dados feita pelo algoritmo da ARIA através da entropia associada definida pela Equação 2.7.

2.3.3 SAINET com raio adaptativo

Tomando por base o que foi dito nas Seções 2.2.2.2 e 2.2.2.3 sobre desempenho e implementação da aiNet e da ARIA,(ALMEIDA, 2017) propôs modificar o algoritmo original da SAINET, substituindo o algoritmo da aiNet pelo da ARIA na construção da população de anticorpos da rede.

A modificação no algoritmo tinha como proposta simplificar a implementação da SAINET, uma vez que o algoritmo da ARIA é mais simples e menos dependente de parâmetros definidos pelo usuário que o da aiNet, e permitir a aplicação da SAINET na classificação de conjuntos de dados cuja distribuição das amostras tenha densidade variável.

O classificador proposto foi testado em duas bases de dados reais, Íris e Vinhos disponíveis em (HETTICH; BLAKE; MERZ, 1998). Os resultados para a base da Iris foram comparados com os resultados apresentados por (KNIDEL, 2006) para modelo original sobre a mesma base, e a SAINET com raio adaptativo apresentou resultados satisfatórios e compatíveis com os resultados da SAINET original. Os resultados da base de Vinhos foram comparados com um outro modelo classificador consolidado, o Naive Bayes (LUCCA et al., 2013), e a acurácia da SAINET com raio adaptativo ficou muito aquém da acurácia do Naive Bayes para a mesma base de dados. Além disso, não houve compressão de dados para a base de vinhos, o que sugere falhas na implementação do algoritmo.

Por fim, (ALMEIDA, 2017) sugere que para melhorar os resultados do algoritmo para base de Vinhos, sejam propostas maneiras de definir o parâmetro r da ARIA, que corresponde ao menor tamanho que o raio mínimo dos anticorpos pode assumir. Também foi sugerido para trabalhos futuros, apresentar uma visualização do posicionamento dos anticorpos em relação aos antígenos que mostre também o raio dos anticorpos para facilitar o entendimento do algoritmo.

3 Propostas para a SAINET com raio adaptativo

Este trabalho se propõem a apresentar duas contribuições principais para o trabalho realizado em (ALMEIDA, 2017):

1. Determinar o parâmetro o raio mínimo dos anticorpos de acordo com a base de dados;

2. Implementar a modificação proposta por (VIOLATO; AZZOLINI; ZUBEN, 2010) para estimar a densidade dos anticorpos no algoritmo da ARIA e verificar o impacto dessa alteração no resultado alcançado pelo classificador

Estas duas propostas de relacionam diretamente com o posicionamento dos anticorpos e, consequentemente, impactam nos resultados do classificador. O raio mínimo r e a densidade do anticorpo são utilizados na Equação 2.2 para calcular o seu raio. Se o valor escolhido para r for inadequado para o conjunto de dados (antígenos), o alcance dos anticorpos é afetado levando a uma representação dos dados que pode não condizer com realidade. Por outro lado, se o método escolhido para determinar a densidade em torno do anticorpo for incongruente, a preservação da distribuição de densidade ao posicionar os anticorpos pode ser prejudicada. Anticorpos mal posicionados em relação aos antígenos criam possibilidade de erro ao classificar novas amostras apresentadas à rede ao testar o modelo, uma vez que a classificação é feita atribuindo à amostra apresentada a classe do anticorpo mais similar a ela (menor distância Euclidiana). Ressalta-se também que o raio dos anticorpos é considerado como limiar de supressão na supressão de rede. Caso um anticorpo esteja localizado dentro da região delimitada pelo raio de outro, significa que estes anticorpos se reconhecem, e portanto aquele que possuir maior raio deve ser eliminado. Assim, é de extrema importância que o raio dos anticorpos seja calculado de forma correta e que sejam condizentes com a base de dados que eles representarão.

3.1 Determinar o raio mínimo r

Observando a Equação 2.2, percebe-se que o raio dos anticorpos será o valor de r multiplicado por um fator que depende da densidade do anticorpo e da dimensão dos dados. Isso implica que quanto menor o valor de r, menores serão os raios do anticorpos que ficarão mais próximos uns dos outros, e quanto maior o valor de r maiores serão os raios dos anticorpos que ficarão mais afastados uns dos outros. A questão é: como determinar se um dado r é muito pequeno ou muito grande para uma determinada base?

Uma possibilidade é olhar para as distâncias entre as amostras da base de dados. Conhecendo as distâncias entre as amostras da base é possível saber o quão perto estão as amostras mais próximas e quão distantes estão as amostras mais afastados. A partir disso, é possível adotar uma métrica para relacionar estas distâncias e o raio mínimo que os anticorpos devem possuir.

Calculando a distância entre todos as amostras da base de dados, é possível montar uma matriz de distâncias que relaciona as amostras e suas distâncias relativas. O raio mínimo r pode ser definido como a média das $S\%$ menores distâncias. Quanto maior for S, maior será r.

Escolher as menores distâncias mantém a localidade do algortimo. As decisões do algoritmo são tomadas observando a vizinhança das amostras. Assim sendo, não se justificaria tomar como parâmetro as maiores distâncias uma vez que elas estão associadas a amostras muito afastados uns dos outros, provavelmente em *clusters* distintos.

3.2 Estimativa da densidade de um anticorpo

O algoritmo original da ARIA determina a densidade de um anticorpo através da contagem dos antígenos localizados na região delimitada pelo raio de vizinhaça E do anticorpo. O parâmetro E é inicializado com um valor aleatório, e atualizado ao final de cada geração com a média aritmética dos raios R de todos os anticporpos que compõem a populaçao atual. Dessa forma, E pode ser maior, menor ou igual ao raio R do anticorpo.

Uma outra maneira de determinar a densidade de um anticorpo é proposta por (VIOLATO; AZZOLINI; ZUBEN, 2010) e (AZZOLINI; VIOLATO; ZUBEN, 2010), es-

timando a densidade em torno de um anticorpo através da Equação 2.8 como feito pelo método KNN. A densidade do anticorpo é inversamente proporcional à distância entre ele e o k-ésimo antígeno mais próximo a ele. Como exemplo, suponha k=3. Se a distância entre x e o terceiro antígeno mais próximo a ele for pequena, a densidade de x será alta, o que indica que os antígenos na vizinhança de x estão muito perto uns dos outros. Por outro lado, se a distância ente **x** e o terceiro antígeno mais próximo a ele for grande, a densidade de x será baixa, indicando que os antígenos na vizinhança de x estão mais afastados uns dos outros. Alternativamente ao que foi proposto por (VIOLATO; AZZOLINI; ZUBEN, 2010) e (AZZOLINI; VIOLATO; ZUBEN, 2010), ao invés de estimar a densidade pela distância para o k-ésimo viznho mais próximos, a densidade foi estimada pela média das distâncias entre o anticorpo e seus k-vizinhos mais próximos. Para efeitos matemáticos, esta alteração não causa grandes alterações no valor da distânica utilizada, mas suaviza a abordagem, fornecendo um ideia mais realista da vizinhança do anticorpo.

4 Estudo de caso

4.1 Bases de Dados

Para realização do estudo de caso, foram escolhidas cinco bases de dados reais, cujos arquivos com informações sobre as instâncias podem ser encontrados no repositório da *University of California, Irvine* (UCI) (HETTICH; BLAKE; MERZ, 1998), para *machine learning*. Além destas, uma base artificial foi gerada para este trabalho, a fim de validar o funcionamento da rede em condições simplificadas e possibilitar a visualização dos dados. A seguir, são apresentados a descrição das bases utilizadas.

1. Iris:

   Uma das bases de dados mais conhecidas na literatura para reconhecimento de padrões, consiste em um conjunto de dados com 150 amostras contendo três classes com 50 instâncias cada. Cada classe corresponde a um tipo de flor Íris: Íris Setosa, Íris Versicolor e Íris Virgínica, sendo que uma delas é linearmente separável das outras duas. A classificação é feita observando os seguintes critérios:

   - Comprimento da sépala em cm
   - Largura da sépala em cm
   - Comprimento da pétala em cm
   - Largura da pétala em cm

2. Vinhos:

   Base de dados composta por um conjunto de 178 resultados de análises químicas realizadas em amostras de vinhos produzidas em uma mesma região da Itália por três produtores distintos. As análises determinam as quantidades de 13 componentes encontrados nas amostras:

   - Álcool

- Ácido málico
- Cinza
- Alcalinidade da cinza
- Magnésio
- Total de fenóis
- Flavonóides
- Fenóis não flavanóides
- Proantocianidinas
- Intensidade de cor
- Hue
- OD280/OD315 de vinhos diluídos
- Proline

Distribuição das classes:

Classe 1: 57 amostras

Classe 2: 71 amostras

Classe 3: 48 amostras

Segundo sua descrição em (HETTICH; BLAKE; MERZ, 1998), a base é considerada bem comportada na separação das classes.

3. Pima Indians Diabetes Database:

   Os dados desta base tem origem em exames clínicos realizados pelo National Institute of Diabetes and Digestive and Kidney Diseases, localizado em Maryland nos Estados Unidos, em pacientes que apresentavam sinais de Diabetes de acordo com os critérios da Organização Mundial de Saúde (OMS). Representa um subconjunto de uma base maior, tendo sido selecionados resultados de pacientes do sexo feminino, com pelo menos 21 anos de idade e que possuíam herança genética da tribo indígena Pima.

Os índios pima, nativos do sudoeste dos Estados Unidos, se tornaram no século 20 o grupo mais obeso e com maior número de diabéticos dos Estados Unidos em decorrência das alterações nos hábitos alimentares aumentando o consumo de gordura e aumento do sedentarismo da população com a integração deste povo à cultura do país (SCILIAR, 2009).

A base utilizada é compos por 768 amostras com 8 atributos, além do atributo de classe:

- Número de gestações
- Concentração de glicose no plasma em 2 horas de teste de tolerância à glicose
- Pressão distólica (mm Hg)
- Espessura da dobra da pelo do tríceps (mm)
- 2 horas em um serum de insulina (mm U/ml)
- Índice de Massa Corporal ((peso em kg/altura em m) elevado ao quadrado)
- Função pedigree diabetes
- idade (anos)
- variável de classe: 1 para positivo para doença ou 0 para negativo para doença

Distribuição das classes:
Classe 1: 268 amostras (postivo para Diabetes)
Classe 0: 500 amostras (negativo para Diabetes)

4. Wisconsin Breast Cancer Database (January 8, 1991):

   Base de dados disponibilizada pelos hospitais da Univerity of Wisconsin - Madison, que contém resultados de exames clínicos realizados em pacientes para detecção de Câncer de mama. As amostras foram recebidas periodicamente à medida que os casos foram reportados pela médico Dr. Woldberg. A base é composta por 699 amostras distribuidas em 8 grupos de acordo com o período em que foram coletadas:

   - Grupo 1: 367 amostras (Janeiro 1989)

- Grupo 2: 70 amostras (Outubro 1989)
- Grupo 3: 31 amostras (Janeiro 1990)
- Grupo 4: 17 amostras (Abril 1990)
- Grupo 5: 48 amostras (Agosto 1990)
- Grupo 6: 49 amostras (Janeiro 1991)
- Grupo 7: 31 amostras (Junho 1991)
- Grupo 8: 86 amostras (Novembro 1991)

Contém 10 atributos, além do atributo de classe:

- Código da amostra
- Espessura do nódulo
- Uniformidade do tamanho da célula
- Uniformidade da forma da célula
- Adesão marginal
- Tamanho de célula epitelial isolada
- Núcleo de Bare
- Cromatina branda
- Nucléolos normais
- Mitoses
- atributo de classe: 2 para benígno ou 4 para malígno

A bases apresenta 16 amostras com atributos faltantes que foram removidos durante o pré-processamento dos dados, reduzindo assim o núemero de amostras para 683 e resultando na seguinte distruição de Distribuição das classes:

Classe 2: 444 amostras

Classe 4: 239 amostras

5. Mk_blobs:

   As amostras desta base de dados foram gerados artificialmente pelo módulo Scikit-learn (PEDREGOSA F. et al., 2011), uma biblioteca do Python com diversas ferramentas para a área de *Machine Learning*. Para isto, foi utilizado o make_blobs, um gerador de amostras que gera amostras isotrópicos com uma distribuição Gaussiana para testar algoritmos de agrupamento e de classificação. Foram geradas 200 amostras formando dois conjuntos distintos, e a cada conjunto foi atribuído um rótulo de classe para que fosse aplicável à SAINET. A visualização das amostras gerados pela ferramenta é exibida na Figura 4.1.

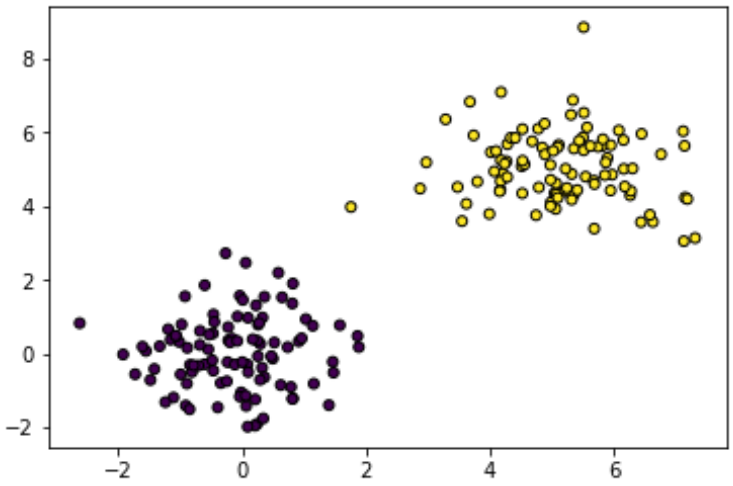

Figura 4.1: Base Mk_blobs gerada artificialmente pela ferramenta make_blobs. Cada conjunto corresponde a uma clase.

As amostras foram geradas com estas características para que fosse possível visualizar antígenos e anticorpos, pois os dados possuem duas dimensões. Além disso, a distribuição dos dados de forma igualitária entre as duas classes que estão muito bem separadas é um cenário ideal para um classificador, reduzindo as chances de erro do modelo.

4.2 Métricas utilizadas na avaliação

Foram adotadas algumas métricas para analisar os resultados alcançados pela rede que serão tratas nas seções seguintes. Nesta seção, são apresentadas suas definições. Para avaliar a acurácia dos modelos classificadores foi é utilizada taxa de classificação correta (TCC), que corresponde à proporção de amostras classificadas de forma correta em relação

ao total de amostras do conjunto de teste. O cálculo da TCC é feito calculando-se o traço da matriz de confusão, que também é utilizada para outras métricas mais adiante. O número de anticorpos (NA) da rede também é exibido nas tabelas deste capítulo. Além destas, a entropia relativa definida pela Equação 2.7 foi calculada depois da primeira fase da SAINET, logo após a construção da população com o algoritmo da ARIA (H inicial) e após a execução da terceira fase da SAINET (H final), com o objetivo de avaliar a variação da entropia relativa entre a população gerada pela ARIA e a população final da rede após o ajuste de pesos.

Além da TCC, são extraídas da matriz de confusão as medidas de sensibilidade (S), especificidade (E) e eficiência, métricas válidas somente para bases de dados com duas classes. A sensibilidade dá a proporção de valores que positivos que de fato foram classificados pelo modelo como positivos através da Equação 4.1. A especificidade por sua vez, dá proporção de valores que negativos que de fato foram classificados pelo modelo como negativos através da Equação 4.2.

$$S = \frac{V_P}{T_P} \tag{4.1}$$

Onde V_P é número acertos positivos e T_P é o total de verdadeiros.

$$E = \frac{V_N}{T_N} \tag{4.2}$$

Onde V_N é número acertos positivos e T_N é o total de verdadeiros.

Especificidade e sensibilidade podem ser relacionadas pela eficiência, que é a média aritmética dos dois valores. Quanto mais próximos de 1 seus valores, melhor é a acurácia do classificador. Para o cálculo de E e S foi adotada a matriz de confusão como segue na Equação 4.3 :

$$MC = \begin{bmatrix} V_N & F_P \\ F_N & V_P \end{bmatrix} \tag{4.3}$$

Algumas tabelas da Seção 4.3 trazem também os valores dos parâmetros r (raio

mínimo) e LRPNP (Limiar de Reconhecimento de Padrões Não-Próprios) calculados.

4.3 Resultados alcançados pelo classificador proposto

Nesta seção são apresentados os resultados obtidos para as alterações propostas para a SAINET com raio adaptativo. Os testes foram realizados seguindo o mesmo padrão para todas das bases. Das amostras disponíveis em cada base de dados, 80% são selecionadas aleatoriamente para treinamento (observando o balanceamento das classes na composição do conjunto de treinamento), e os 20% restantes para teste. A divisão do conjunto de dados foi feita desta forma para que Os resultados gerados estivesse em conformidade com (ALMEIDA, 2017) e (KNIDEL, 2006) para comparações posteriores. Os resultados apresentados são referentes aos valores médios obtidos em 30 execuções do algoritmo com os parâmetros indicados.Em todos os testes foram adotados =0,3 (exceto para base da Iris, em que foi adotado =0,1) e μ=0,95, e número de gerações igual a 30. A Tabela 4.1 apresenta os resultados obtidos na execução do algoritmo da SAINET com raio adaptativo usando a proposta de definição do raio mínimo dos anticorpos como a média entre as 20% menores distâncias entre os elementos da base, e com o cálculo da densidade feito pelo algoritmo original da ARIA.

A Tabela 4.1 apresenta os resultados obtidos pela SAINET para as bases de dados utilizadas adotando r e LRPNP calculados e a densidade dos anticporpos determinada pelo algoritmo original da ARIA. Nela são exibidos além dos valores dos parâmetros calculados, o valores obtidos da entropia relativa, TCC e NA.

Base de dados	r	LRPNP	H inicial	H final	TCC	NA
Iris	0,06	0,07	0,38±0,06	0,73±0,06	73,11±12,53	7,03±0,51
Índios Pima	0.28	0.34	0,67±0,03	0,80±0,09	53,74±15,15	21,2±1,57
Vinhos	0,014	0,017	0,26±0,07	0,33±0,09	70,26±4,8	16,16±1,56
BC_Wisconsin	0,37	0,45	-2,09±0,18	-2,17±0,18	84,57±12,31	166,86±5,66
Mk_blobs	0,14	0,18	1,81±0,18	1,780±0,11	100,0±0,0	6,32±0,82

Tabela 4.1: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e cálculo original da densidade. Valores obtidos para o raio mínimo r, limiar de reconhecimento de padrões não-próprios (LRPNP), e valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Bases de dados normalizadas através da L2-Norm.

Os valores obtidos de sensibilidade, especificidade e eficiência para as bases de

dados com duas classe são exibidos na Tabela 4.2.

Base de dados	E	S	Eficiência
Ìndios Pima	0,56±0,41	0,53±0,43	0,54±0,42
BC_Wisconsin	0,86±0,17	0,86±0,19	0,85±0,18
Mk_blobs	1,0±0,0	1,0±0,0	1,0±0,0

Tabela 4.2: Resultados complementares para as bases de dados com duas classes da aplicação da SAINET com raio adaptativo, com raio mínimo calculado e cálculo original da densidade. Valores médios obtidos para Especificidade (E), sensibilidade (S) e eficiência.

A Tabela 4.1 sugere que conjuntos de dados diferentes requerem valores diferentes de r, e que portanto usar um mesmo valor de r para duas bases de dados distintas pode não ser adequado. De fato, esta afirmação faz sentido uma vez que os valores dos atributos das amostras das bases de dados podem apresentar faixas diferentes de valores, além da distribuição dos dados que também pode não ser semelhante. Segundo (KNIDEL, 2006) não existe um valor de LRPNP universal, desta forma, este parâmetro também deve ser definido de acordo com a base de dados. Para os testes aqui realizados, o valor de LRPNP é definido em função do r calculado, tendo sido fixado em $1,2 \times r$. Este valor foi apenas uma sugestão. Os testes apresentados por (KNIDEL, 2006) sugerem que ora o melhor valor de LRPNP é maior, ora é menor que r, variando de uma base para outra. Apenas para a título de complemento, aqui foram testadas duas possibilidades de LRPNP $= 0,8 \times r$ e LRPNP $= 1,2 \times r$, não apresentando diferenças significantes estatisticamente. Por este motivo, uma delas foi escolhida por preferência e estes resultados não serão colocados aqui.

A mesma ideia se aplica à taxa de aprendizado α. Para as bases onde era possível gerar uma visualização do conjunto de dados (Iris e Mk_blobs), alguns valores foram atribuídos para α (0,05; 0,1; 0,3; 0,5) observando o comportamento refletido no posicionamento dos anticorpos na rede e na taxa de classificação. O valor de $\alpha = 0,1$ foi escolhido para Iris por ter apresentado melhores resultados. Para a base Mk_blobs, o melhor valor foi $\alpha = 0,3$. Para as demais, só era possível comparar a variação nos resultados do classificador, e novamente não houve diferença significativa nos resultados. Desta forma, o valor de $\alpha = 0,3$ foi também aplicado aos demais conjuntos de dados.

As Tabelas 4.3 a 4.7 apresentam os resultados obtidos na classificação quando substituindo o cálculo da densidade original no algoritmo da ARIA pela estimativa feita

através da Equação 2.8 para cada uma das bases de dados. Para esta avaliação, foram escolhidos valores k a serem comparados a fim de encontrar o k mais adequado para cada base. De forma análoga ao que foi feito para as bases com duas classes no algoritmo original, as tabelas 4.4, 4.6 e 4.7 apresentam também os valores de especificidade, sensibilidade e eficiência obtidos para cada valor de k para as referidas bases.

K	H inicial	H final	TCC	NA
3	0,82±0,06	1,01±0,22	62,67±7,32	3,93±0,25
5	0,99±0,11	1,24±0,28	66,56±11,14	2,77±0,62
7	1,00±0,09	1,38±0,29	61,88±10,87	2,77±2,76
10	0,90±0,04	1,06±0,11	71,11±7,90	2,91±0,04

Tabela 4.3: Base Iris: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Parâmetros adotados: r=0,06, LRPNP=0,07, =0,1. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm.

K	H inicial	H final	TCC	NA	E	S	Eficiência
3	1,88±0,08	1,45±0,87	71,36±1,66	6,37±0,87	0,53±0,20	0,8±0,20	0,67±0,21
5	0,93±0,05	1,28±0,17	72,70±9,21	6,27±0,51	0,51±0,13	0,85±0,18	0,68±0,16
7	1,04±0,04	1,45±0,18	67,47±3,17	5,33±0,60	0,53±0,14	0,75±0,11	0,68±0,15
10	0,92±0,07	1,12±0,21	69,78±3,75	7,87±0,99	0,55±0,09	0,78±0,08	0,34±0,08

Tabela 4.4: Base Índios Pima: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,28, LRPNP=0,34, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm.

K	H inicial	H final	TCC	NA
3	0,94±0,12	1,24±0,21	58,64±0,50	4,32±0,64
5	0,72±0,12	1,21±0,23	70,27±9,31	4,8±0,4
7	0,77±0,01	1,15±0,17	76,13±6,31	5,0±0,0
10	0,90±0,00	0,95±0,12	60,18±5,02	4,0±0,26

Tabela 4.5: Base Vinhos: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), e número de anticorpos (NA). Parâmetros adotados: r=0,014, LRPNP=0,017, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm.

K	H inicial	H final	TCC	NA	E	S	Eficiência
3	0,27±0,35	3,08±0,96	64,96±0,0	1±0,0	0,0±0,0	1,0±0,0	0,50±0,0
5	0,13±0,07	3,27±1,07	64,96±0,0	1±0,0	0,0±0,0	1,0±0,0	0,50±0,0
7	0,19±0,23	2,97±0,95	64,96±0,0	1±0,0	0,0±0,0	1,0±0,0	0,50±0,0
10	0,03±0,11	1,67±0,24	64,96±0,0	1±0,0	0,0±0,0	1,0±0,0	0,50±0,0

Tabela 4.6: Base BC_Wisconsin: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,37, LRPNP=0,45, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm.

K	H inicial	H final	TCC	NA	E	S	Eficiência
3	1,81±0,18	1,78±0,11	99,91±0,44	2,45±0,80	1,0±0,0	1,0±0,0	1,00±0,0
5	1,09±0,29	1,58±0,15	100,00±0,0	3,83±1,02	1,0±0,0	1,0±0,0	1,00±0,0
7	0,97±0,00	1,42±0,02	100,00±0,0	5,0±0,0	1,0±0,0	1,0±0,0	1,00±0,0
10	0,94±0,13	1,37±0,01	100,00±0,0	4,8±0,9	1,0±0,0	1,0±0,0	1,00±0,0

Tabela 4.7: Base Mk_Blobs: Resultados da aplicação da SAINET com raio adaptativo com raio mínimo calculado e estimativa da densidade para valores k (k = 3, 5, 7 e 10). Valores médios obtidos para entropias relativas (H final e H inicial), taxa de classificação correta (TCC), número de anticorpos (NA), especificidade (E), sensibilidade (S) e eficiência. Parâmetros adotados: r=0,14, LRPNP=0,18, =0,3. Os valores de r e LRPNP foram os mesmos calculados para os testes cujos resultados estão descritos na Tabela 4.1. Bases de dados normalizadas através da L2-Norm.

Os resultados mostrados nas Tabelas 4.3 a 4.8 indicam que, de forma análoga a r e LRPNP, também não há um valor de k universal. Valores diferentes de k foram observados no melhor resultado, considerando TCC e NA. Para as bases dos Índios Pima e Mk_blobs, o melhor k observado foi k=3. Para esta última, k=3 foi escolhido por ter sido o que apresentou menor número de anticorpos na rede. Para a base da Iris, temos k=7, para a base de Vinhos, temos k=7. Para BC_Wisconsin, todos os valores testados de k geraram resultados equivalentes.

Além da acurácia, é necessário também avaliar a taxa de compressão da rede para as bases de dados. As Tabelas 4.8 a 4.13 indicam as taxas de compressão atingidas pela rede usando o cálculo orginal da densidade e usando a densidade estimada. A taxa de compressão foi calculada sobre o tamanho do conjunto efetivamente utlizado para treinamento, uma vez que o tamanho do conjunto utilizado poder menor que 80% do total de amostras devido ao balanceamento das classes. O cálculo da taxa de compressão considera os tamanhos médios dos conjuntos de treinamento e a quantidade média de anticorpos das 30 execuções do algoritmo da SAINET feitas durante os testes.

Base	Conjunto treinamento	Número de anticorpos	Taxa de Compressão
Iris	113,14±2,94	7,03±0,51	93,79%
Índios	432,2±12,11	21,2±1,57	95,09%
Vinhos	109,5±18,45	16,16±1,56	85,24%
BC_Wisconsin	381,6±9,06	166,86±5,66	56,27%
Mk_blobs	157,5±3,32	6,32±0,82	98,99%

Tabela 4.8: Taxas de compressão médias obtidas para as bases de dados em relação ao conjunto de treinamento utilizado adotando o cálculo original da densidade no algoritmo da ARIA. Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

k	Conjunto treinamento	Número de anticorpos	Taxa de Compressão
3	114,0±10,28	3,93±0,25	96,55%
5	114,32±11,29	2,73±0,4	97,60%
7	113,5±12,45	2,77±0,62	97,56%
10	114,10±11,06	2,91±0,04	97,45%

Tabela 4.9: Taxas de compressão médias obtidas para a base da Iris em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

k	Conjunto treinamento	Número de anticorpos	Taxa de Compressão
3	431,0±13,0	6,37±0,82	98,52%
5	432,1±12,29	6,27±0,51	98,55%
7	430,99±11,45	5,33±0,60	98,76%
10	431,03±11,83	7,87±0,99	98,17%

Tabela 4.10: Taxas de compressão médias obtidas para a base dos Índios Pima em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

k	Conjunto de treinamento	Número de anticorpos	Taxa de Compressão
3	110,1±12,0	4,32±0,64	96,08%
5	109,4±11,09	4,8±0,4	95,61%
7	109,7±11,45	5,0±0,0	95,44%
10	108,93±10,83	4,2±0,26	96,1%

Tabela 4.11: Taxas de compressão médias obtidas para a base de Vinhos em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

k	Conjunto de treinamento	Número de anticorpos	Taxa de Compressão
3	386,4±9,24	1,0±0,0	99,74%
5	384,8±10,1	1,0±0,0	99,74%
7	386,2±9,85	1,0±0,0	99,74%
10	985,93±10	1,0±0,0	99,74%

Tabela 4.12: Taxas de compressão médias obtidas para a base BC_Wisconsin relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

k	Conjunto de treinamento	Número de anticorpos	Taxa de Compressão
3	157,8±3,23	2,45±0,8	98,44%
5	158,2±2,99	3,83±1,2	97,57%
7	157,9,2±3,65	5,0±0,0	96,83%
10	157,4±3,1	4,8±0,9	96,95%

Tabela 4.13: Taxas de compressão médias obtidas para a base Mk_Blobs em relação ao conjunto de treinamento utilizado adotando a estimativa da densidade no algoritmo da ARIA (valores de k adotados: 3, 5, 7 e 10). Tamanhos médios do conjuntos de treinamentos de cada base considerando que todas as classes possuíam o mesmo número de amostras no conjunto, número de anticorpos (NA) médio e taxa de compressão média.

4.4 Comparações da acurácia

Nesta seção, os resultados alcançados pelo modelo com as novas propostas são comparados com resultados obtidos por outros classificadores da literatura para as bases de dados aqui analisadas.

4.4.1 Comparação com a SAINET com raio adaptativo

A Tabela 4.14 compara os resultados obtidos pela SAINET com raio adaptativo proposta por (ALMEIDA, 2017) (versão original) com os resultados obtidos neste trabalho (versão modificada).

	Versão original		Versão modificada			
			Original		k	
	TCC	NA	TCC	NA	TCC	NA
Iris	93,0±2,8	16,3±2,45	73,11±12,53	7,03±0,51	71,11±7,9	2,91±0,04
Vinhos	92,33±5,38	142,0±0,0	70,26±4,8	16,16±5,66	76,13±6,31	5,0±0,0

Tabela 4.14: Comparação entre os resultados obtidos para as bases da Iris e de Vinhos com os resultados obtidos por (ALMEIDA, 2017) para as mesmas bases de dados. Valores comparados: valores médios da taxa de classificação correta (TCC) e do número de anticorpos (NA).

A Tabela 4.15 compara os parâmetros usados em cada abordagem. Em (ALMEIDA, 2017), as bases de dados não foram normalizadas. Nesta tabela, são mostrados os valore de r e LRPNP calculados para as bases com e sem a normalização dos dados.

	Sem normalização		Sem normalização		Com normalização	
	r	LRPNP	r	LRPNP	r	LRPNP
Iris	0,5	0,8	0,82	0,98	0,06	0,07
Vinhos	0,5	0,4	78,57	94,28	0,014	0,017

Tabela 4.15: Valores de r e LRPNP adotados neste trabalho em comparação com os utilizados por (ALMEIDA, 2017), onde os dados não foram normalizados. Comparados com os valores calculados para tais parâmetros considerando as bases sem normalização e com normalização L2-Norm.

Os valores adotados na versão original fazem parte dos valores usados por (KNIDEL, 2006) em seus testes, tendo sida a combinação que apresentou melhor resultado. A Tabela 4.15 sugere que ainda para dados normalizados, pode ser difícil definir um vlaor de r. Para a base de Vinhos em especial, embora a taxa de classificação correta média em (ALMEIDA, 2017) tenha sido superior ao alcançado com a nova abordagem, não houve

compressão de dados. Para este caso, calcular o valor de r e de LRPNP como proposto uma vez que o número de anticorpos foi reduzido drasticamente de uma abordagem para a outra.

Sobre as demais bases de dados, por não haver referências na literatura sobre a aplicação da SAINET com raio adaptativo a elas, não há como comparar a estratégia de cálculo de r. Por este motivo, a acurácia da SAINET sobre estas bases de dados foi comparada a de um outro classificador como apresentado a seguir.

4.4.2 Comparação com Naive Bayes

A fim de avaliar a acurácia (TCC) da SAINET com as modificações sugeridas, foram realizados testes com outros classificadores conhecidos na literatura. A Tabela 4.16 traz uma comparação dos resultados médios de 30 execuções do algoritmos Naive Bayes (NB) (LUCCA et al., 2013). A base Mk_blobs não foi testada para este modelo.

Base de dados	TCC
Iris	93,10±0,0
Indios Pima	100,0 ±
Vinhos	71,43±0,0
BC_Wisconsin	87,5±0,0

Tabela 4.16: Resultados da classifcação usando Gaussian Naive Bayes.

Para as bases da Iris e dos Índios Pima, o NB apresentou acurácia superior ao das implementações da SAINET com raio adaptativo. Para a base de Vinhos, o NB teve acurácia superior à SAINET usando o cálculo original da densidade na estimativa usando k=3 e k=10, equivalente à implementação usando k=5, e inferior à implementação com k=7. Para a base BC_Wisconsin, O NB apresentou acurácia muito próxima à da SAINET usando o cálculo original da densidade, e superior à todos os valores de k.

Não foram realizados testes para a base Mk_blobs usando o Naive Bayes em decorrência de um problema com a conversão do dados para o formato esperado pelo algoritmo para a leitura. Contudo, a ausência desta comparação não prejudica os resultados obtidos pela rede para esta base uma vez que a acurácia alcançada foi satisfatória em todas as abordagens analisadas.

5 Conclusões e Trabalhos Futuros

Apesar da proposta do algoritmo da ARIA de preservar a distribuição de densidade do conjunto original de dados ao representá-lo por protótipos que recebem o nome de anticorpos, um estudo teórico aponta que a ARIA não é capaz de preservar a distribuição de densidade em bases com altas dimensões. A partir desta ideia, foi questionado se isto afetaria os resultados da SAINET com raio adaptativo, visto que esta utiliza o algoritmo da ARIA pra construção da população de anticorpos. Para tanto, uma modificação sugerida pela literatura na forma como a densidade dos anticorpos é estimada pela ARIA foi implementada, estimando-se a densidade na vizinhança do anticorpo considerando a distância média do anticorpo em questão para seus k antígenos mais próximos.

Os resultados dessa avaliação sugerem que a estimativa da densidade aumenta a taxa de compressão, ou seja, diminui o número de anticorpos da rede, em relação ao cálculo original. Uma possível justificativa para isto é que a estimativa não leva em conta o raio de vizinhança E, definido como a média dos raios R dos anticorpos, que por sua vez é calculado em função da densidade dos anticorpos. Esta queda significativa do número de anticorpos pode explicar a queda da acurácia em alguns casos, como da Iris e BC_Wisconsin.

Sobre a acurácia da rede classificadora, para algumas bases foi possível encontrar um valor de k que melhorasse a acurácia da rede. Foi o caso das bases de Vinhos e dos Índios Pima, tendo tido esta última o maior incremento da acurácia. A meta de melhorar os resultados da rede sobre a base de Vinhos aumentando a taxa de compressão sugerida por (ALMEIDA, 2017) foi alcançada, uma vez que houve compressão dos dados da forma esperada e apesar de a taxa de classificação correta ter sido inferior à apresentada anteriormente, estimando-se a densidade com k=7, a acurácia da SAINET com raio adaptativo supera a do NB.

Sobre a proposta de cálculo do raio mínimo r e de relacioná-lo com o LRPNP, ainda são necessários mais estudos para chegar uma conclusão definitiva. Contudo, os resultados sobe as bases de Vinhos e BC_Wisconsin da rede comparados com os do NB

sugerem que o valor calculado de r foi acertado.

Quando analisamos os valores da entropia relativa calculada pela Equação 2.7, ao contrário do que era esperado , não se observa grandes mudanças entre cálculo original da densidade e a estimativa de densidade. É provável que isto se dê pelo fato de, para as bases testadas, a entropia relativa usando a ARIA original já apresentar valores baixos (quanto mais próximos de zero, melhor) o que não deixa margem para grandes reduções. Ao que tudo indica, a qualidade do posicionamento dos anticorpos varia muito de uma abordagem para outra para os casos testados.

Em linhas gerais, podemos dizer que o posicionamento dos anticorpos isoladamente não influência tanto no resultado da classificação. Durante a fase de construção da rede a informação de classe dos anticorpos não é considerada. Considerando apenas a distância entre os anticorpos na supressão de rede, é possível que sejam eliminados todos os anticorpos de uma classe em regiões onde as classes não estão bem separadas. Isto resulta em um posicionamento coerente dos anticorpos do ponto de vista do agrupamento, mas leva a erros na classificação já que uma classe do conjunto de dados não estaria representada na rede. Para evitar que esses casos aconteçam, principalmente quando a densidade for estimada, seria necessário garantir que todas as classes estejam representadas na população.

Quanto aos trabalhos futuros, sugere-se considerar as classes dos anticorpos na supressão de rede para evitar a perda de representação das classes. Além disto, é possível também utilizar outras divisões para os conjuntos de treinamento e teste em bases de dados com poucos amostras (como por exemplo: 70% para treinamento e 30% para teste, ou 60% para treinamento e 40% para teste) uma vez que nestes casos ter poucas amostras para testar modelo pode prejudicar os resultados tendo em vista que cada amostra representa uma porcentagem grande conjunto de teste quando este possuir poucas amostras.

Pode-se também utilizar outras abordagens baseadas em grafos, como conjunto dominante por exemplo, para encontrar um valor adequado para os raios dos anticorpos e também para r e o LRPNP levando em conta a distribuição dos dados a fim de eliminar a dependência do ajuste do raio de parâmetros definidos pelo usuário, o que pode melhorar a acurácia do modelo.

Uma outra proposta, é aplicar a SAINET com raio adaptativo à bases de dados com grande número de atributos e de amostras para avaliar a compressão da rede e seu desempenho em bases de dados com alta dimensão. Além disto, podem ser utilizadas bases de dados com outros tipos de atributos (que não sejam numéricos) e aplicar também outros cálculos de distância além da Euclidiana, como por exemplo a distância Manhattan ou qualquer outra distância que seja adequada ao tipo dos atributos.

Bibliografia

ABBAS A. K., L. A. H. P. S. Imunologia celular e molecular. In: *Imunologia Celular e Molecular*. [S.l.]: Elsevier, 2009. p. 4–6.

ALMEIDA, A. L. S. S. Redes imunológicas artificiais para classificação: Sainet com raio adaptativo. In: . [S.l.]: Universidade Federal de Juiz de Fora, 2017.

AZZOLINI, A. G.; VIOLATO, R. P.; ZUBEN, F. J. V. Density preservations and vector quantization in immune-inspired algorithm. In: . [S.l.: s.n.], 2010. p. 36–46.

BEURER, G. M. Análise de dados de altas diemnsões. In: . [S.l.]: Universidade Federal do Rio Grande de Sul, 2010.

BEZERRA, G. et al. Adaptive radius immune algorithm for data clustering. In: *Adaptive radius immune algorithm for data clustering*. [S.l.: s.n.], 2005. p. 290–303.

CAMILO, C.; SILVA, J. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. In: . [S.l.: s.n.], 2009.

CASTRO, L. N. de; ZUBEN, F. J. V. Artificial immune systems: Part i – basic theory and applications. In: . [S.l.: s.n.], 1999.

CASTRO, L. N. de; ZUBEN, F. J. V. ainet: An artificial immune network for data analysis. In: . [S.l.]: Idea Group Publishing, 2002. p. 231–259.

DARWIN, C. A origem das espécies. In: . [S.l.: s.n.], 1859.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. In: . [S.l.]: American Association for Artificial Intelligence, 2010.

FIGUEREDO, G.; BERNARDINO, H.; BARBOSA, H. Introdução aos sistemas imunológicos artificiais. In: *Introdução aos Sistemas Imunológicos Artificiais*. [S.l.: s.n.], 2013.

FRANçA, F. O. de et al. Conceptual and pratical aspects of the ainet family of algorithms. In: *Conceptual and Pratical Aspects of the aiNet Family of Algorithms*. [S.l.: s.n.], 2010. p. 1–35.

HETTICH, S.; BLAKE, C. L.; MERZ, C. J. Uci repository of machine learning databases. In: . [S.l.: s.n.], 1998.

JERNE, N. K. Towards a network theory of the immune system. In: *Towards a network theory of the immune system*. [S.l.: s.n.], 1974. p. 378–389.

KNIDEL, H. Extensões e aplicações de redes neuro-imunológicas. In: . [S.l.]: Universidade Federal de Campinas, 2006. p. 39–41,62–65.

KOHONEN, T. . Self-organizing maps. In: *Self-organizing maps*. [S.l.]: Springer, 2010.

LOPES, G. A. de M. imarag: Um algoritmo baseado em sistemas imunológicos artificiais para treinamento de redes mlp. In: *iMaRag: Um algoritmo baseado em Sistemas Imunológicos Artificiais para treinamento de redes MLP*. [S.l.]: Universidade de Pernambuco, 2010.

LUCCA, G. et al. Uma implementação do algoritmo naïve bayes para classificação de texto. In: . [S.l.: s.n.], 2013.

PARHAM, P. The immune system. In: *Imunologia Celular e Molecular*. [S.l.]: Garland Science, 2009. p. 4–7.

PEDREGOSA F., V. G. G. A. M. V. T. B. et al. Scikit-learn: Machine learning in Python. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2011. p. 2825–2830, publisher = , edition = , journal = Journal of Machine Learning Research, key = Pedregosa et.al, 2011, location = United States.

POOLE, D.; MACKWORTH, A. Artificial intelligence: foundations of computational agents. In: . [S.l.]: Cambridge University Press, 2010.

SCILIAR, M. Diabetes: Aprendendo com os índios. In: . [S.l.]: Academia Brasileira de Letras, 2009.

STIBOR, T.; TIMMIS, J. An investigation on the compression quality of ainet. In: *Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*. [S.l.: s.n.], 2007. p. 495–501.

VIOLATO, R. P. V.; AZZOLINI, A. G.; ZUBEN, F. J. V. Antibodies with adaptative radius as prototypes of high-dimensional datasets. In: *Lecture Notes in Computer Science (LNCS) 6209*. [S.l.: s.n.], 2010. p. 158–170.

ZUBEN, F. J. V. Computação evolutiva: Uma abordagem pragmática. In: . [S.l.]: Unicamp, 2011.