



Universidade Estadual de Feira De Santana
Programa de Pós-Graduação em Computação Aplicada

AILINE - Um Método Baseado em Redes Neurais Artificiais para Detecção Automática de Linhas Espectrais em Galáxias na Região do Óptico

Yvson Paulo Nascimento Ferreira

Feira de Santana

2017



Universidade Estadual de Feira De Santana
Programa de Pós-Graduação em Computação Aplicada

Yvson Paulo Nascimento Ferreira

**AILINE – Um Método Baseado em Redes Neurais
Artificiais para Detecção Automática de Linhas
Espectrais na Região do Óptico**

Dissertação apresentada à Universidade
Estadual de Feira de Santana como parte
dos requisitos para a obtenção do título de
Mestre em Computação Aplicada.

Orientador: Prof. Dr. Iranderly Fernandes de Fernandes

Coorientador: Prof. Dr. Angelo Conrado Loula

Feira de Santana

2017

Ficha Catalográfica - Biblioteca Central Julieta Carteado

F444a Ferreira, Yvson Paulo Nascimento
AILINE - Um método baseado em redes neurais artificiais para
detecção automática de linhas espectrais em galáxias na região do óptico /
Yvson Paulo Nascimento Ferreira. - 2017.
108 f.: il.

Orientador: Iranderly Fernandes de Fernandes.

Coorientador: Angelo Conrado Loula.

Dissertação (mestrado) - Universidade Estadual de Feira de
Santana, Programa de Pós-Graduação em Computação Aplicada, 2017.

1. Redes neurais (Computação). 2. Espectroscopia. I. Fernandes,
Iranderly Fernandes de, orient. II. Loula, Angelo Conrado, coorient.. III.
Universidade Estadual de Feira de Santana. IV. Título.

CDU: 004.8

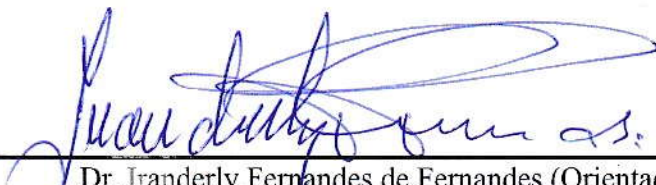
Yvson Paulo Nascimento Ferreira

**AILINE - Um Método Baseado em Redes Neurais Artificiais para
Detecção Automática de Linhas Espectrais em Galáxias na Região
do Óptico**

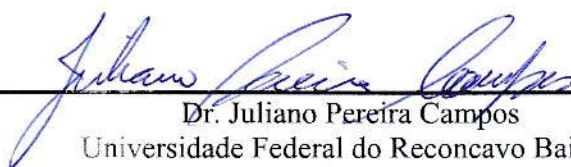
Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 24 de agosto de 2017

BANCA EXAMINADORA



Dr. Iranderly Fernandes de Fernandes (Orientador)
Universidade Estadual de Feira de Santana



Dr. Juliano Pereira Campos
Universidade Federal do Recôncavo Baiano



Dr. Eduardo Brescansin de Amôres
Universidade Estadual de Feira de Santana

Abstract

Advances in the acquisition technology of astronomical spectra resulted in an enormous amount of data. Not being more feasible to analyze them using classical approaches, the need for automatic methods arises. Then, in this research is presented, an Intelligent Algorithm for Identifying Spectral Lines, the AILINE (in Portuguese), which utilizes an artificial neural network to identify the emission lines in the optical spectra of galaxies. This method that in the tests carried out has achieved a accuracy higher than 95% is evaluated and faced with other automatic approaches and other machine learning algorithms.

Keywords: Artificial Neural Nets (ANN), astroinformatics, astronomy, emission lines, machine learning, spectroscopy.

Resumo

Os avanços na tecnologia de aquisição de espectros astronômicos resultaram em uma enorme quantidade de dados. Não sendo mais viável analisá-los usando abordagens clássicas, surge a necessidade de métodos automáticos. Então, nesta pesquisa é apresentado um Algoritmo Inteligente para Identificação de Linhas Espectrais, o AILINE, que utiliza uma Rede Neural Artificial para identificar as linhas em emissão nos espectros ópticos de galáxias. Este método que nos testes realizados alcançou uma acurácia superior a 95%, é avaliado e confrontado com outras abordagens automáticas e outros algoritmos de aprendizado de máquina.

Palavras-chave: aprendizado de máquina, astroinformática, astronomia, espectroscopia, linhas em emissão, Redes Neurais Artificiais (RNA).

Prefácio

Esta dissertação de mestrado foi submetida a Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. Iranderly Fernandes de Fernandes e coorientador o Dr. Angelo Conrado Loula.

Agradecimentos

Em primeiro lugar, agradeço a Deus, pois sem Ele nada seria.

Agradeço aos meus avós e pais pela criação e educação proporcionada, sem as quais não teria chegado aqui.

Agradeço a compreensão de minha esposa por ter sido paciente e atenciosa durante estes anos em que fui mais ausente para poder me dedicar a este projeto.

Ao meu orientador, Prof. Dr. Iranderly Fernandes e ao meu coorientador Prof. Dr. Angelo Loula, agradeço pela paciência e compreensão. Agradeço pelas aulas, reuniões e todo conhecimento que me foi passado. Sem este auxílio, este trabalho não seria possível.

A minha família e amigos, os meus sinceros agradecimentos pelo carinho e atenção dedicados todo esse tempo.

Sumário

Abstract	i
Resumo	ii
Prefácio	iii
Agradecimentos	iv
Sumário	v
Lista de Publicações	vii
Lista de Tabelas	viii
Lista de Figuras	xi
Lista de Abreviações	xiii
Lista de Símbolos	xiv
1. Introdução	1
1.1 Objetivos	3
1.2 Contribuições.....	4
1.3 Organização do Trabalho.....	5
2. Fundamentação	6
2.1 Astroinformática.....	6
2.2 Espectroscopia	7
2.3 Trabalhos Relacionados	10
2.4 Redes Neurais Artificiais.....	13
3. Definição do Problema	16
3.1 Delimitações	17
4. Metodologia	19
4.1 Plataforma de Desenvolvimento	21
4.2 Arquitetura da Rede Neural Artificial.....	21
4.3 Base de Dados	22
4.4 Pré-Processamento.....	24
4.5 Método e Medidas de Avaliação	26
4.6 Experimentos.....	31

5. Resultados e Discussões	37
5.1 Treinamento e Testes de Validação.....	37
5.2 Treino e Testes Finais	64
5.3 Consistência Externa.....	70
5.4 Comparação com o ALFA.....	75
6. AILINE	76
6.1 Tempo de Execução.....	79
6.2 <i>AILINE-Training</i>	79
7. Considerações Finais	81
7.1 Aplicações.....	82
7.2 Pesquisas Futuras.....	83
8. Referências Bibliográficas	84
9. Espectros Utilizados na Pesquisa	88

Lista de Publicações

- Uma Proposta para Identificação Automática de Linhas em Emissão em Espectros Ópticos de Galáxias. Yvson Ferreira, Iranderly Fernandes, Angelo Loula. (2016). *In WPOS/ERBASE*.
- AILINE-Um Método Inteligente para Detecção Automática de Linhas Espectrais em Galáxias. Yvson P. N. Ferreira, Iranderly F. de Fernandes, Angelo C. Loula. (2017). *In KDMILE/SBBD*.

Lista de Tabelas

Tabela 4.1: Resumo dos espectros que compõem a base de dados agrupados por galáxias.....	23
Tabela 4.2: Resumo dos dados de treino e testes.....	26
Tabela 5.1: Resultados do treinamento com os parâmetros do bloco de experimentos 1	38
Tabela 5.2: Valores estatísticos do teste de Friedman para o recall do bloco de experimentos 1.....	38
Tabela 5.3: Valores estatísticos do teste de Friedman para a precisão do bloco de experimentos 1.....	39
Tabela 5.4: Valores estatísticos do método de Bonferroni para a precisão do bloco de experimentos 1.....	40
Tabela 5.5: Valores estatísticos do teste de Friedman para a acurácia do bloco de experimentos 1.....	41
Tabela 5.6: Valores do método de Bonferroni para a acurácia do bloco de experimentos 1.....	42
Tabela 5.7: Valores estatísticos do teste de Friedman para a AUC do bloco de experimentos 1.....	43
Tabela 5.8: Valores do método de Bonferroni para a AUC do bloco de experimentos 1.	44
Tabela 5.9: Medidas de avaliação para os experimentos do bloco 2.....	46
Tabela 5.10: Valores estatísticos do teste de Friedman para o recall do bloco de experimentos 2.....	47
Tabela 5.11: Valores do método de Bonferroni para o recall do bloco de experimentos 2	47
Tabela 5.12: Valores estatísticos do teste de Friedman para a precisão do bloco de experimentos 2.....	48
Tabela 5.13: Valores do teste de Bonferroni para a precisão do bloco de experimentos 2	49
Tabela 5.14: Valores estatísticos do teste de Friedman para a acurácia do bloco de experimentos 2.....	50
Tabela 5.15: Valores do método de Bonferroni para a acurácia do bloco de experimentos 2.....	51

Tabela 5.16: Valores estatísticos do teste de Friedman para a AUC do bloco de experimentos 2.....	52
Tabela 5.17: Valores do método de Bonferroni para a AUC do bloco de experimentos 2	53
Tabela 5.18: Medidas de avaliação dos parâmetros para o bloco de experimentos 3..	55
Tabela 5.19: Valores estatísticos do teste de Friedman para o recall do bloco de experimentos 3.....	56
Tabela 5.20: Valores estatísticos do método de Bonferroni para o recall do bloco de experimentos 3.....	57
Tabela 5.21: Valores estatísticos do teste de Friedman para a precisão do Bloco de experimentos 3.....	58
Tabela 5.22: Valores do método de Bonferroni para a precisão do Bloco de experimentos 3.....	59
Tabela 5.23: Valores estatísticos do teste de Friedman para a acurácia do bloco de experimentos 3.....	60
Tabela 5.24: Valores do método de Bonferroni para a acurácia do bloco de experimentos 3.....	61
Tabela 5.25: Valores estatísticos do teste de Friedman para a AUC do bloco de experimentos 3.....	62
Tabela 5.26: Valores do método de Bonferroni para a AUC do bloco de experimentos 3	63
Tabela 5.27: Medidas de avaliação para as amostras de teste T1 e T2.....	65
Tabela 5.28: Medidas de avaliação para as amostras de teste T2 e T3.....	65
Tabela 5.29: Matriz de confusão para amostra de teste T2	66
Tabela 5.30: Matriz de confusão para amostra de teste T3	66
Tabela 5.31: Comparação de medidas dos classificadores testados para duas amostras de testes.....	70
Tabela 5.32: Valores estatísticos do teste de Friedman para o recall dos classificadores	71
Tabela 5.33: Valores estatísticos do teste de Friedman para a precisão dos classificadores	71
Tabela 5.34: Valores estatísticos do método de Bonferroni para a precisão dos classificadores	71
Tabela 5.35: Valores estatísticos do teste de Friedman para a acurácia dos classificadores	72
Tabela 5.36: Valores estatísticos do método de Bonferroni para a acurácia dos classificadores	73
Tabela 5.37: Valores estatísticos do teste de Friedman para a AUC dos classificadores	73

Tabela 5.38: Valores estatísticos do método de Bonferroni para a AUC dos classificadores	74
Tabela 6.1: Exemplo da tabela gerada pelo AILINE com identificação dos íons	78

Lista de Figuras

Figura 1.1 Intervalos da radiação eletromagnética. Fonte: Site Knoow.net.....	3
Figura 2.1: Espectro da galáxia IIZW 107.....	8
Figura 2.2: Representação esquemática da absorção de fótons a partir da transição de elétrons entre níveis atômicos.....	9
Figura 2.3: Representação esquemática da emissão de fótons a partir da transição de elétrons entre níveis atômicos.....	9
Figura 2.4: Modelo básico para um nó. x_i =entrada, W_i =Peso, f =função de transferência, y =saída	14
Figura 2.5: Representação de uma Rede Neural Artificial com múltiplas camadas....	14
Figura 3.1: Espectro da galáxia MRK 309 com destaque de linhas fortes e fracas.....	17
Figura 4.1: Linha em emissão 4959.52[OIII] em três espectros de galáxias distintas..	20
Figura 4.2: Topologia básica da rede neural artificial.....	22
Figura 4.3: Exemplo de um pico de linha em emissão	25
Figura 4.4: Exemplo de linhas fortes, fracas e ruído	27
Figura 4.5: Exemplo de uma Curva ROC.....	30
Figura 4.6: Processo de parada por checagens de validação	34
Figura 4.7: Variedade de soluções no espaço de soluções.....	35
Figura 5.1: Gráfico para o teste de Bonferroni para a precisão do bloco de experimentos 1.....	41
Figura 5.2: Gráfico do método de Bonferroni para a acurácia do bloco de experimentos 1.....	43
Figura 5.3: Gráfico do método de Bonferroni para a AUC do bloco de experimentos 1	45
Figura 5.4: Gráfico do método de Bonferroni para o recall do bloco de experimentos 2	48
Figura 5.5: Gráfico do método de Bonferroni para a precisão do bloco de experimentos 2.....	50
Figura 5.6: Gráfico do método de Bonferroni para a acurácia do bloco de experimentos 2.....	52
Figura 5.7: Gráfico do método de Bonferroni para a AUC do bloco de experimentos 2	

.....	54
Figura 5.8: Gráfico do método de Bonferroni para o recall do bloco de experimentos 3	56
.....	58
Figura 5.9: Gráfico do método de Bonferroni para a precisão do bloco de experimentos 3.....	60
Figura 5.10: Gráfico do método de Bonferroni para a acurácia do bloco de experimentos 3.....	62
Figura 5.11: Gráfico do método de Bonferroni para a AUC do bloco de experimentos 3	67
Figura 5.12: Curva ROC para a amostra de teste T2.....	67
Figura 5.13: Curva ROC para a amostra de teste T3.....	68
Figura 5.14: Espectro da galáxia MRK 712 com classificações da RNA	68
Figura 5.15: Espectro da galáxia NGC 4385 com classificações da RNA	69
Figura 5.16: Sobreposição de linhas espectrais de três galáxias	72
Figura 5.17: Gráfico do método de Bonferroni para a precisão dos classificadores	73
Figura 5.18: Gráfico do método de Bonferroni para a acurácia dos classificadores	74
Figura 5.19: Gráfico do método de Bonferroni para a AUC dos classificadores.....	77
Figura 6.1: Exemplo de gaussiana ajustada em um pico.....	77
Figura 6.2: Exemplo de ajuste do continuum	78
Figura 6.3: Exemplo do gráfico gerado pelo AILINE com identificação dos íons.....	

Lista de Abreviações

Abreviação	Descrição
AILINE	Algoritmo Inteligente para Identificação de Linhas Espectrais
ALFA	<i>Automated Line Fitting Algorithm</i>
AT	Algoritmos Testados
AUC	<i>Area Under the Curve</i>
BRB	<i>Bayesian Regularization Backpropagation</i>
CCD	<i>Charge-Coupled Device</i>
FN	<i>False Negative</i>
GAMA	<i>Galaxy And Mass Assembly</i>
LAMOST	<i>Large sky Area Multi-Object Spectroscopic Telescope</i>
LMB	<i>Levenberg-Marquardt Backpropagation</i>
MATLAB	<i>MATrix LABoratory</i>
MRK	Galáxias <i>Markarian</i>
MLP	<i>Multilayer Perceptron</i>
NGC	Galáxias do <i>New General Catalog</i>
NTT	<i>New Technology Telescope</i>
PNA	Plano Nacional de Astronomia
QN	Quantidade de Neurônios
RNA	Rede Neural Artificial
ROC	<i>Receiver Operating Characteristic</i>
SDSS	<i>Sloan Digital Sky Survey</i>
SVM	<i>Support Vector Machine</i>
TN	<i>True Negative</i>
TP	<i>True Positive</i>
ZW	Galáxias <i>Zwicky</i>

Lista de Símbolos

Símbolo	Descrição
Å	Angstrom (Unidade de medida de comprimento. $1\text{Å} = 10^{-10}\text{m}$.)
χ^2	Valor da distribuição estatística para o ajuste de <i>Bonferroni</i>
H_0	Hipótese Nula

Capítulo 1

Introdução

“A gravidade explica os movimentos dos planetas, mas não pode explicar quem colocou os planetas em movimento. Deus governa todas as coisas e sabe tudo que é ou que pode ser feito”

-- Isaac Newton

A Astronomia busca contar a história do universo, e ao longo de sua existência sempre se beneficiou das últimas novidades tecnológicas e teorias científicas. E nessa busca procura responder, ou ajudar a responder, perguntas como: Estamos sozinhos no Universo? Existem outros planetas semelhantes à Terra? Se o *Big Bang* aconteceu como seus destroços deram origem às estrelas e planetas? E como a vida surgiu de um início cósmico sem vida?

Existem cientistas que esperam ter algumas dessas respostas em breve, devido às descobertas da Astronomia nas últimas décadas, como por exemplo, a observação da expansão acelerada do universo, o mapeamento das primeiras ondulações da matéria, a identificação de milhares de planetas fora do Sistema Solar e ainda devido às informações obtidas e esperadas por meio dos projetos científicos em andamento e daqueles por vir (DALCANTON *et al.*, 2015).

Como exemplos desses projetos, podem ser citados os atuais mapeamentos do Universo chamados de *mega-surveys*, que são capazes de coletar informações de milhares de objetos celestes em uma noite. Dentre os principais destes estão o *Sloan Digital Sky Survey* (SDSS), que até a sua décima segunda *data release* já tinha a

imagem, em cinco bandas, de mais de um terço da esfera celeste e obtido mais de cinco milhões de espectros astronômicos (ALAM *et al.*, 2015). E o *Large sky Area Multi-Object Spectroscopic Telescope* (LAMOST) que em sua primeira *data release* já obteve mais de dois milhões de espectros astronômicos (LUO *et al.*, 2015).

Esses e os demais projetos proporcionam uma verdadeira avalanche de dados que já ultrapassam a capacidade humana para análise manual, o que tem proporcionado o surgimento de novas abordagens para lidar com estes dados, como a Astroinformática e os Observatórios Virtuais que viabilizam formas mais dinâmicas para armazenamento, tratamento, distribuição e análise dos dados por meio do desenvolvimento de rotinas automatizadas, protocolos e metodologias que estão muitas das vezes na atual fronteira do conhecimento (BORNE, 2009).

Uma das principais fontes de informações dos objetos astronômicos é a espectroscopia, que estuda a interação da radiação eletromagnética com a matéria e cujas técnicas são as melhores ferramentas para a investigação e análise químico-física em áreas tão diversas quanto a engenharia de alimentos, indústria petroquímica, biomedicina e astrofísica (REQUENA *et al.*, 2007). Suas aplicações na Astronomia vão desde as medições dos campos magnéticos na superfície do Sol até a derivação da composição química e características físicas de galáxias distantes (VON BERLEPSCH, 2011).

O resultado gráfico de uma observação por espectroscopia é chamado de espectro, e na Astronomia o espectro eletromagnético é estudado em praticamente toda a sua extensão a partir dos raios gama, passando pelos Raios-x, ultravioleta, visível, infravermelho até o rádio, sendo que cada região espectral tem suas peculiaridades e por isso cada uma demanda uma tecnologia própria (SALA, 2008). A região do espectro primeiramente estudada foi a região do visível que compreende o intervalo de 350 a 750 nm (Figura 1.1) e atualmente seu estudo é conhecido como espectroscopia óptica, a qual por muitos anos foi a maior fonte de informações para a Astronomia, sendo ainda hoje a faixa do espectro eletromagnético de grande importância para o entendimento do universo (KITCHIN, 1995).

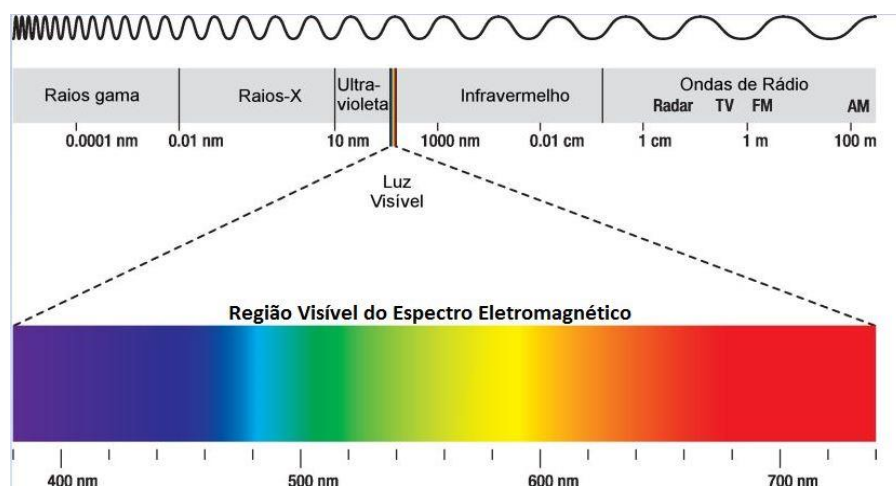


Figura 1.1 Intervalos da radiação eletromagnética. Fonte: sítio Knowow.net¹.

Em um espectro podem ser observadas as linhas em emissão, absorção e o contínuo, que surgem devido às transições dos elétrons nas órbitas dos átomos. Identificar e medir as linhas espectrais é um processo crucial para a análise do espectro, pois é a partir desta identificação que se estimam as informações do objeto observado (KITCHIN, 1995). A análise destas linhas acaba, em certo ponto, se distinguindo uma da outra pelas particularidades apresentadas por elas.

Então, devido ao atual montante de dados coletados na Astronomia e à importância da análise destes, faz-se necessário a criação ou aprimoramento dos métodos para tal análise, de forma a proporcionar aos pesquisadores a capacidade de realizá-la tão rapidamente quanto os dados são produzidos.

1.1 Objetivos

Devido à importância da identificação das linhas espectrais no processo de análise dos dados espectroscópicos e os problemas em aberto nesse campo, o objetivo principal desta pesquisa é desenvolver uma abordagem para identificação automática de linhas em emissão nos espectros de galáxias na faixa do óptico, de forma individual ou em lote, com níveis de sinal/ruído e resolução espectral variados, com o mínimo de intervenção humana possível, e que seja de fácil aplicação.

¹Disponível em: <http://knowow.net/cienciasexactas/fisica/espectroelectromagnetico/>. Acesso em 16/07/2017.

Os objetivos específicos são:

- Revisar as abordagens existentes para identificação de linhas espectrais;
- Treinar algoritmos de aprendizado de máquina para classificar as linhas em emissão no espectro;
- Avaliar a eficácia da proposta desenvolvida;
- Possibilitar o treino de Redes Neurais Artificiais para classificação de linhas em emissão.

1.2 Contribuições

O Plano Nacional de Astronomia (PNA) expressa os anseios e necessidades da comunidade científica brasileira ligada à Astronomia. Neste documento constam alguns dos objetivos a serem alcançados por essa comunidade nos próximos anos, dentre eles está uma participação mais efetiva do país no avanço do conhecimento ligado a várias vertentes da Astronomia.

Como o desenvolvimento de novos instrumentos representa um investimento econômico significativo, neste plano reconhece-se que uma área em que o Brasil poderia contribuir mais efetivamente seria no desenvolvimento de softwares (PLANO NACIONAL DE ASTRONOMIA, 2010).

Este reconhecimento é válido, principalmente pela necessidade crescente de rotinas automáticas para viabilizar a análise do atual volume de dados coletado pela Astronomia, inclusive na forma de dados provindos da espectroscopia óptica, cuja análise demanda bastante rigor visto a importância das informações contidas neste tipo de dado para uma melhor compreensão do universo.

Mas, apesar das necessidades esboçadas, muito desta análise ainda é feita de forma manual ou semiautomática, e muitos dos sistemas existentes atualmente para esses fins são monousuários, de difícil manuseio e instalação e que exigem uma grande curva de aprendizado para seu correto uso.

Assim sendo, entre as contribuições possíveis nesta área está, o desenvolvimento de metodologias confiáveis, que permitam identificar e medir as linhas em emissão dos espectros automaticamente, pois, esta identificação permanece

entre as principais etapas da análise de espectros, e ainda hoje apresenta desafios a serem superados.

Pelos pontos abordados anteriormente pode ser entendido que o desenvolvimento de rotinas automáticas para análise espectral, e que possam suprir as necessidades apontadas, estão dentro dos anseios da comunidade científica do país e do mundo. Neste trabalho propõe-se uma abordagem para esse fim, que seja mais efetiva do que a análise manual e de fácil manuseio para poder ser usada por especialistas experientes, bem como iniciantes.

Uma contribuição adicional além do desenvolvimento da rotina computacional, é a verificação da viabilidade do uso de redes neurais para classificação de linhas espectrais.

1.3 Organização do Trabalho

Esta pesquisa foi dividida em sete capítulos, contando com esta introdução.

No Capítulo 2 é apresentada a fundamentação teórica sobre os temas abordados no trabalho: Astroinformática, Espectroscopia, Revisão das metodologias utilizadas para a identificação de linhas em emissão e Rede Neurais artificiais. No Capítulo 3 será detalhado o problema chave que esta pesquisa procura resolver e as delimitações da mesma. No Capítulo 4 será apresentada a metodologia utilizada desde a plataforma de desenvolvimento escolhida até os detalhes sobre os experimentos que serão realizados. No Capítulo 5 serão apresentados e discutidos os resultados dos testes realizados para a escolha do método de treinamento da rede neural artificial e as avaliações desta rede em comparação a outros classificadores e a outras abordagens para identificação de linhas espectrais. No Capítulo 6 a rotina computacional desenvolvida que recebeu o nome de AILINE será exposta. E finalmente no Capítulo 7 serão discutidas as considerações finais sobre este trabalho, bem como as conclusões obtidas e as aplicações e sugestões para pesquisas futuras.

Capítulo 2

Fundamentação

“O que sabemos é uma gota; o que ignoramos é um oceano”

-- Isaac Newton

Neste capítulo estão sendo revistos os conceitos básicos de que tratam esta pesquisa, bem como as últimas pesquisas que visam resolver ou minimizar os problemas aqui destacados.

2.1 Astroinformática

Com o aumento da capacidade de coletar dados diversas áreas de pesquisa estão armazenando-os em uma velocidade muito maior do que podem analisá-los, sendo a Astronomia uma dessas áreas, principalmente com o avanço dos grandes mapeamentos celestes como o SDSS (YORK *et al.*, 2000) e o LAMOST (ZHAO *et al.*, 2006), dentre outros.

O SDSS é um dos principais levantamentos do universo da atualidade, segmentado em vários projetos² como: APOGEE, eBOSS, MaNGA, dentre outros. E desde 1998 já proveu informações sobre milhões de objetos astronômicos por meio de seu telescópio exclusivo de 2,5m localizado no *Apache Point Observatory* (Novo México, EUA), equipado com uma câmera CCD (*Charge-Coupled Device*) que fotografa o céu em 5 bandas fotométricas (u , g , r , i e z). Inicialmente contava com dois espectrógrafos multifibras que podiam observar 640 objetos simultaneamente,

² Para mais detalhes sobre o SDSS ver: <http://www.sdss.org/>.

porém, atualizações são feitas e novos equipamentos são adicionados. Até sua décima segunda *Data Release*, finalizada em 2015, um total de 470 milhões de objetos em imagens e 5,3 milhões de espectros foram disponibilizados em seus catálogos (ALAM *et al.*, 2015).

O LAMOST é um grande levantamento que utiliza um telescópio de 4m, localizado no Observatório de *Xinglong (Beijing, China)*, que possui em seu plano focal quatro mil fibras para aquisição de espectros em alta velocidade. E desde 2008 está dedicado a um levantamento espectroscópico dos objetos celestes do Hemisfério Norte. Em 2012 disponibilizou os dados de seu projeto piloto, contando com 717.660 espectros, contendo 648.820 estrelas, 2.723 galáxias e 621 quasares. Em sua primeira *Data Release* em 2015 já tinha atingindo a marca de 2.204.860 espectros de galáxias, quasares e estrelas (LUO *et al.*, 2015).

Levantamentos do céu como esses elevaram o volume de dados de *gigabytes* a *terabytes* em poucos anos, e rapidamente será alcançada a casa dos *petabytes* devido aos novos projetos cada vez mais ambiciosos. Com tantos dados cada vez mais os pesquisadores desta área necessitam de apoio da informática para realizarem suas análises. Surge então uma nova disciplina que está sendo chamada por alguns de Astroinformática (BORNE, 2009), que se baseia tanto na informática como nos avanços da análise estatística (BROMOVÁ *et al.*, 2014).

A Astroinformática assume um caráter interdisciplinar entre a informática e a Astronomia, trazendo da informática técnicas ligadas à organização e descrição de dados, mineração de dados e aprendizado de máquina, além de toda a infraestrutura necessária para armazenar, disponibilizar e processar esses dados. Já da Astronomia herda as taxonomias de classificação astronômicas, ontologias e conceitos, juntamente com todo *know-how* desta ciência milenar (BORNE, 2009).

2.2 Espectroscopia

Uma das áreas da Astronomia que tem se beneficiado diretamente dessa avalanche de dados é a espectroscopia, que estuda as interações da matéria (sólido, líquido ou gás) com a radiação ao longo de todo espectro eletromagnético provendo ricos detalhes sobre os processos físicos e químicos que ocorrem nas galáxias e demais

objetos estudados, como: medir a distâncias das galáxias, determinar sua composição química e verificar a natureza de suas populações estelares, verificar temperatura pressão e gravidade, dentre outras (HOPKINS *et al.*, 2013).

Ela não se beneficia apenas do aumento do volume de dados, mas também da qualidade dos dados espectroscópicos que estão sendo coletados, permitindo novas descobertas sobre os objetos estudados e o entendimento do universo como um todo. Como a exemplo da descoberta e estudo de exoplanetas orbitando estrelas distantes, que se beneficia da qualidade dos espectros de alta definição obtido graças aos modernos instrumentos e às novas abordagens científicas. Espectros de alta precisão permitem ainda entre outras coisas, verificar a hidrodinâmica da atmosfera estelar e resolver estruturas de linhas em absorção intergalácticas em quasares (DRAVINS, 2010).

Vale salientar que a espectroscopia não é uma técnica somente usada na Astronomia, mas é amplamente utilizada em várias áreas da ciência, como biomedicina, engenharia de alimentos, bioquímica, biofísica, dentre outras. Sendo assim, as melhorias nessa técnica permitem um avanço em todas as áreas que a utilizam (SCHOLKMANN *et al.*, 2012).

As informações obtidas por meio da espectroscopia advêm da análise do espectro, que pode ser visualizado por um gráfico de intensidade das linhas em função do seu respectivo comprimento de onda Figura 2.1, a qual apresenta um espectro da galáxia IIZW 107.

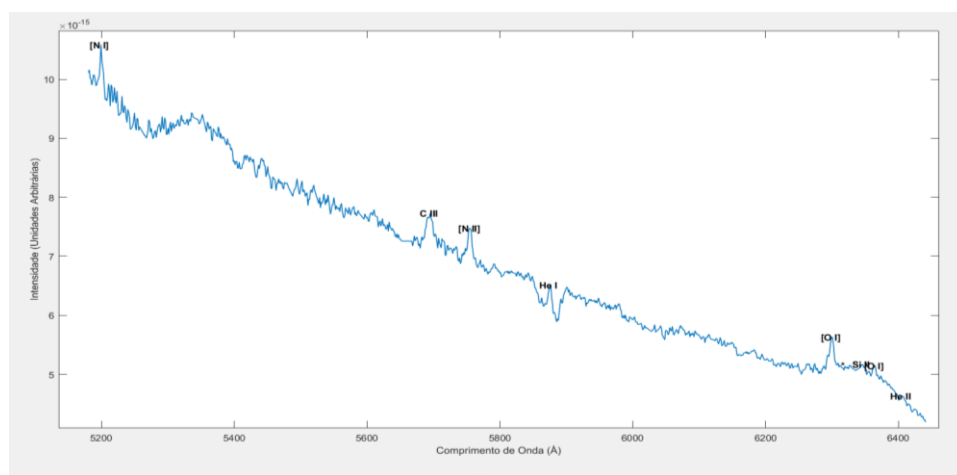


Figura 2.1: Espectro da galáxia IIZW 107.

Sempre que a quantidade de energia em um átomo muda, ocorre o surgimento de uma linha espectral, quando um fóton com energia suficiente é absorvido o átomo pode ser excitado ou ionizado, de maneira que um de seus elétrons salta para outro nível de energia mais excitado, formando uma linha de absorção (Figura 2.2), quando o elétron salta para um nível de energia menos excitado, o fóton é emitido, formando assim uma linha em emissão (Figura 2.3) (HETEM e PEREIRA, 2010).

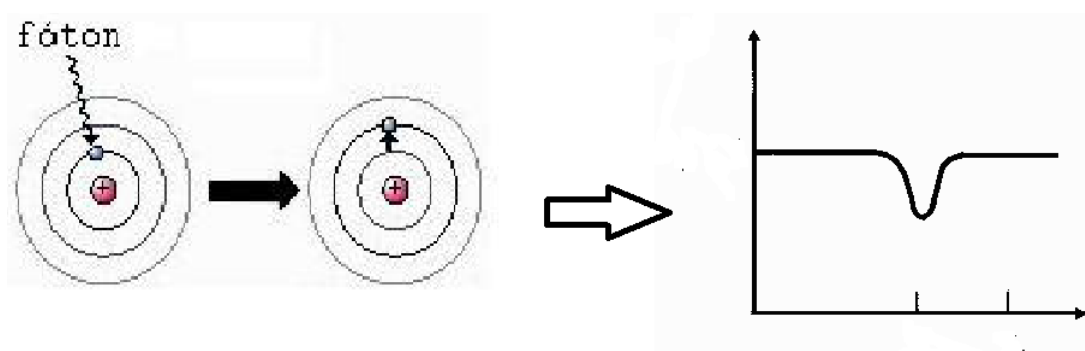


Figura 2.2: Representação esquemática da absorção de fótons a partir da transição de elétrons entre níveis atômicos.

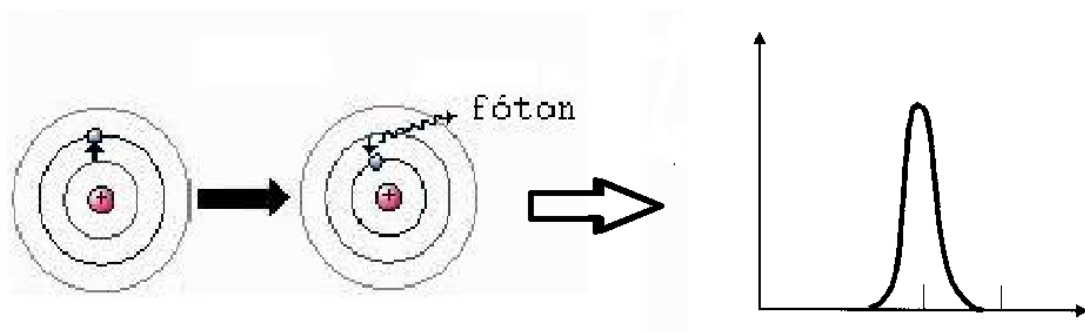


Figura 2.3: Representação esquemática da emissão de fótons a partir da transição de elétrons entre níveis atômicos.

As linhas em emissão podem ser encontradas na maioria dos objetos astronômicos estudados e ao longo de todo espectro eletromagnético, sendo muito úteis para verificar as características químico-físicas destes objetos³.

Para uma correta interpretação dos dados espectroscópicos se faz necessária uma precisa análise das linhas espectrais. Pois, é por meio da investigação destas

³ Para uma visão mais ampla sobre as informações obtidas por meio das linhas em emissão ver: STASINSKA, 2007. Disponível on-line em: <https://arxiv.org/pdf/0704.0348.pdf>. Acesso em 10/10/2017.

linhas que se obtêm uma grande variedade de informações. Esta análise começa com a identificação das linhas, que é um processo minucioso e que exige atenção e experiência (WESSON, 2016).

Kitchin (1995) descreve os passos da abordagem tradicional para esta identificação que começa com a distinção do que representa ruído e o que efetivamente é uma linha espectral. Depois o comprimento de onda de todas as possíveis linhas deve ser medido para que a partir de um catálogo, ou outra fonte, se identifique as linhas encontradas. Com isto pode-se estimar a velocidade radial do objeto, e corrigir todos os comprimentos de onda das linhas aos seus valores de repouso.

Com o uso de tabelas é possível encontrar os átomos e íons correspondentes aos do espectro. Neste ponto pode ser feito uma verificação no espectro das linhas identificadas com algum grau de incerteza para as que deveriam existir em determinado átomo ou íon. Então começa um processo iterativo de ajustes do que os catálogos e tabelas indicam em relação às linhas encontradas, até o máximo de linhas possíveis ser identificado com o maior grau de precisão possível.

Como esta identificação depende de comparação visual, e por ser um processo exaustivo, principalmente para espectros de alta resolução que contam com um grande número de linhas, acaba sendo suscetível a erros e dependente do grau de experiência do especialista. Mesmo em uma abordagem semiautomática esse processo pode ter erros e demandar bastante tempo (WESSON, 2016).

Algo notável e que merece ser relatado, é que mesmo nas abordagens mais atuais, com processos semiautomáticos ou automáticos, onde para uma mesma amostra de espectros, especialistas diferentes usando as mesmas rotinas, ou usando rotinas diferentes, buscando identificar as linhas em emissão, acabam encontrando resultados divergentes quanto às linhas identificadas (WESSON, 2016).

2.3 Trabalhos Relacionados

Apesar da existência de várias rotinas semiautomáticas implementadas em reconhecidas ferramentas para análise espectral, como IRAF, CLASS, DIPSO, XSpec,

CASA, GANDALF, SPLAT, SHERPA, MIDAS e outras, ainda existe a necessidade de uma solução padrão que minimize, ou de preferência elimine, as discrepâncias citadas anteriormente e reduza o tempo de análise. O termo semiautomático é utilizado porque para que as linhas sejam reconhecidas é necessário seguir uma série de passos para parametrizar a rotina afim de utilizá-la. Assim, apesar de serem um avanço em relação à análise manual, as mesmas ainda colocam uma grande carga sobre os usuários e a identificação final ainda fica dependente da experiência e atenção do usuário. Dessa forma, o tempo de análise com essas rotinas pode levar de dias a meses para o atual volume de dados coletados em apenas uma noite.

Um passo importante para uma melhor solução é a descoberta precisa e automática dos picos das linhas em emissão. E como a descoberta de picos em sinais é um passo importante no processamento de sinais, ao longo do tempo vários métodos foram desenvolvidos para esse fim, como os baseados em janela de limiar (VIVÓ-TRUYOLS *et al.*, 2005), transformada *wavelet* (DU *et al.*, 2006), técnicas usando *templates* (MTETWA *et al.*, 2006), filtragem não linear (SHIM *et al.*, 2009), filtragem usando segunda derivada gaussiana (FREDRIKSSON *et al.*, 2009), e estatística de alta ordem (PANOULAS *et al.*, 2001).

Embora esses algoritmos funcionem bem em situações específicas, eles apresentam deficiências para aplicações mais gerais, necessitando de ajustes de parâmetros a depender do uso, ou a depender da taxa de sinal-ruído, o que acaba demandando tempo e introduzindo erros no processo (SCHOLKMANN *et al.*, 2012).

As opções automáticas existentes estão em sua maioria introduzidas em *pipelines* usados pelos grandes levantamentos como o SDSS e o *Galaxy And Mass Assembly* (GAMA) descrito em Hopkins *et al.* (2013) e outros trabalhos promissores como os de Hong *et al.* (2014). Estes utilizam técnicas como ajustes iterativos de *templates*, que se baseiam em estimativas estatísticas para identificar e medir as gaussianas ajustadas às linhas esperadas. No entanto, ainda são carentes de muitas melhorias, pois, como pode ser visto em Hopkins *et al.* (2013), muitas classificações são apontadas com incertezas muito grandes, e apesar dos vários mecanismos de redundância para as classificações, ainda em muitos casos é necessário a verificação final por especialistas humanos. E no caso de Hong *et al.* (2014), apenas a

identificação de linhas fortes é enfatizado.

No entanto, foi encontrada uma proposta mais recente e mais próxima do proposto neste trabalho, ou seja, permitir localizar e classificar os picos de linhas em emissão com precisão, de forma automática, e com o mínimo de interferência humana, chamada de ALFA (*Automated Line Fitting Algorithm*) (WESSON, 2016).

Nesta abordagem um algoritmo genético é usado para otimizar os parâmetros necessários para gerar espectros sintéticos para comparação com o espectro observado. Para isto, o ALFA primeiramente estima quais linhas em emissão poderiam estar presentes no espectro, em seguida constrói um espectro sintético de perfis gaussianos para cada linha estimada, um algoritmo genético é utilizado para otimizar os parâmetros das funções gaussianas utilizadas. Desta forma, na primeira geração são criados por volta de 30 espectros, é então calculada a qualidade do ajuste por meio da soma do erro quadrático entre o espectro sintético e o original, o melhor membro da população é escolhido para fazer parte da próxima geração e os 30% melhores servem de base para a próxima geração, estes vão sofrer mutações, conforme os parâmetros definidos para formar esta nova geração. O processo é repetido por 500 gerações e o melhor membro da última geração é escolhido como o espectro que será utilizado.

Apesar de representar um grande avanço, como essa abordagem usa um algoritmo genético e é muito dependente da sua estimativa inicial dos espectros, quando esta estimativa é ruim, os resultados finais também o são, além de que no final o espectro real é ignorado e as medições são feitas no espectro simulado, sem contar que os resultados mudam toda vez que se verifica o mesmo espectro. Nos testes aqui realizados para verificar sua eficácia, com os mesmos espectros utilizados neste trabalho, foi observado que em uma estimativa haviam duas linhas identificadas pelo ALFA, e em outra ele estimou oito, já em um teste realizado com espectros fornecidos junto com o código, a estimativa variou em dezenas de linhas de um teste para o outro no mesmo espectro. Portanto, resultados como estes apontam a necessidade de um refinamento nos parâmetros da abordagem.

Contudo, o uso de algoritmos de aprendizado de máquina, como as redes neurais artificiais têm alcançado resultados promissores em identificar padrões

diversos em dados espectroscópicos adquiridos em observações astronômicas e em outras áreas afins, como em Tu *et al.* (2008), em que uma rede foi treinada para classificar estrelas de acordo com os seus tipos principais. E ainda para localizar picos em sinais de radar (WUNSCH, 2015) e em sinais de eletrocardiograma (VIJAYA *et al.*, 1998). Em ambos os trabalhos foi reportado que as redes neurais artificiais conseguiram assimilar o padrão dos picos e classificá-los com uma boa precisão.

Não foi encontrado na literatura nenhum trabalho que utilize Redes Neurais Artificiais (a partir de agora chamada de RNAs) para a identificação exclusiva de todas as linhas em emissão em espectros de galáxias da forma pretendida neste trabalho, porém seu uso para classificar determinados padrões de picos nos espectros como os trabalhos de Wunsch (2015) e Vijaya *et al.* (1998), e ainda trabalhos como o de Tu *et al.* (2008) onde foram utilizadas RNAs para classificações de estrelas por meio de seus espectros, indicam que isso seja possível.

De acordo ao que foi visto na literatura, é de suma importância o desenvolvimento e/ou aprimoramento de abordagens automáticas para a identificação e análise das linhas espectrais, que permitam não apenas cobrir todo o volume de dados atualmente disponível, em tempo hábil, o que é urgentemente necessário para os dados astronômicos (ZHANG *et al.*, 2009), como também permita uma melhor precisão, o que é possível devido ao caráter imparcial da máquina em relação a subjetividade humana frente às dificuldades inerentes dessa análise (BYKOV, 2004).

2.4 Redes Neurais Artificiais

O uso de algoritmos de aprendizado de máquina e, sobretudo, de redes neurais artificiais na classificação e análise de dados espectrais é uma tendência crescente e como nesta pesquisa esta abordagem será utilizada se faz necessário uma descrição dos principais aspectos desta técnica.

As RNAs são uma categoria de algoritmos de aprendizado de máquina que possuem sua inspiração nas redes neurais biológicas, como por exemplo, o cérebro humano; tais redes são compostas por neurônios mais simples que se interconectam para executar tarefas complexas e diversificadas.

Em uma RNA o modelo básico de um neurônio artificial ou nó, como é conhecido, recebe múltiplas entradas que são associadas a pesos e então é feita uma soma ponderada cujo resultado passa o sinal por meio de uma função de transferência para o próximo nó (LIVINGSTONE, 2008). A Figura 2.4 ilustra este processo.

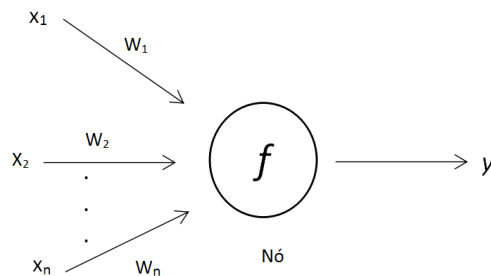


Figura 2.4: Modelo básico para um nó. x_i =entrada, W_i =Peso, f =função de transferência, y =saída.

Porém geralmente um neurônio só não é suficiente, por isso são utilizados vários neurônios interligados em uma camada operando em paralelo, ou em camadas diversas, neste modelo a saída de uma camada é a entrada de outra, a conexão de vários neurônios em uma ou mais camadas é chamada RNA (DEMUTH *et al.*, 2014). Uma ilustração de uma RNA pode ser vista na Figura 2.5.

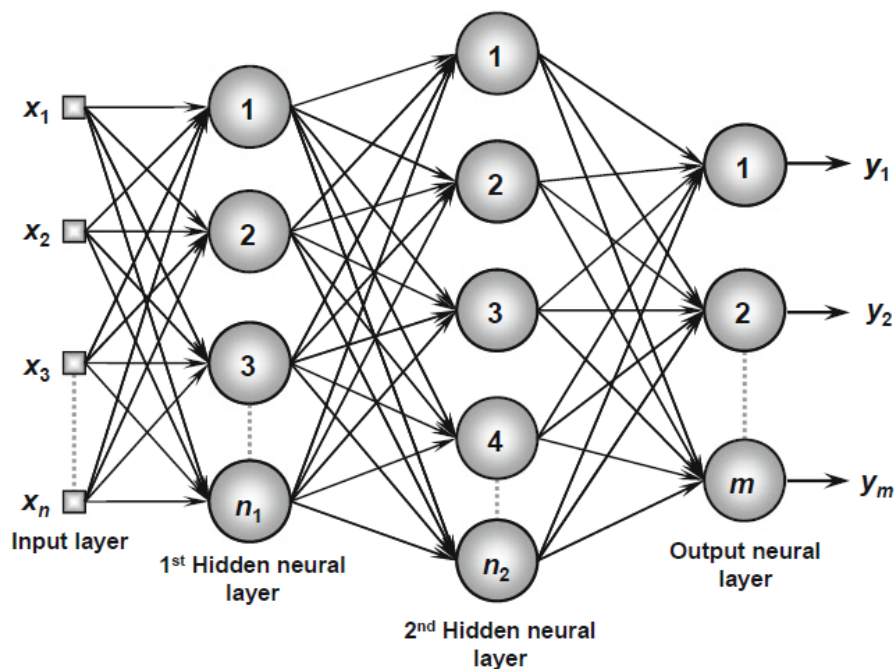


Figura 2.5: Representação de uma Rede Neural Artificial com múltiplas camadas. Fonte: (DA SILVA *et al.*, 2016).

O aprendizado da RNA basicamente se dá pelo ajuste dos pesos durante o processo de treinamento, que pode ser realizado de dois modos: supervisionado ou não supervisionado. No aprendizado supervisionado um conjunto de exemplos, que são compostos pelas entradas associadas às saídas desejadas, é fornecido e os pesos são ajustados de forma a minimizar o erro entre a saída da rede e as saídas desejadas fornecidas. Já no aprendizado não supervisionado, este conjunto de exemplos não é fornecido e então a rede procura descobrir sozinha o padrão ou tendência existente a partir da entrada fornecida (LIVINGSTONE, 2008).

A RNA mais comumente conhecida, e uma das primeiras a ser utilizada, é a *Multilayer Perceptron* (MLP), cujo treinamento normalmente é feito por meio de um algoritmo de retro propagação (*backpropagation*) onde a primeira entrada é propagada através da rede para que a saída seja calculada por meio de uma função de custo (o erro calculado entre a saída obtida e a saída desejada). Estes valores são enviados de volta a entrada para que os pesos sejam ajustados, este processo acontece iterativamente em várias épocas de treino, e a cada época busca-se diminuir o erro (LIVINGSTONE, 2008). E nesta pesquisa será utilizada a RNA do tipo MLP que utiliza o aprendizado supervisionado.

Capítulo 3

Definição do Problema

“Todas as verdades são fáceis de entender uma vez que são descobertas; o ponto é descobri-las.”

-- Galileu Galilei

A análise dos espectros na Astronomia contribui muito para a determinação de diversas propriedades químicas e físicas dos objetos astronômicos observados (HOPKINS *et al.*, 2013). No entanto, devido às peculiaridades das linhas espectrais, sua correta classificação e análise, quando feita de forma manual, se torna dependente da capacidade do especialista, o que além de ser passível de erros, pode levar muito tempo a depender da resolução e quantidade de espectros a analisar (BYKOV, 2004).

Quando se visualiza um espectro como o exibido na Figura 3.1, é possível verificar que as linhas em emissão quando muito fortes se destacam do ruído, porém, quando muito fracas, acabam sendo confundidas com o próprio ruído, e dependendo da resolução espectral é comum ter linhas combinadas. Então a correta classificação demanda tempo e muita atenção do especialista, somando-se a isto, nos tempos atuais o avanço das tecnologias que tem permitido a realização de grandes levantamentos espectroscópicos elevam a quantidade de dados a serem analisados a um número acima da capacidade humana de análise (YORK *et al.*, 2000; ZHAO *et al.*, 2006).

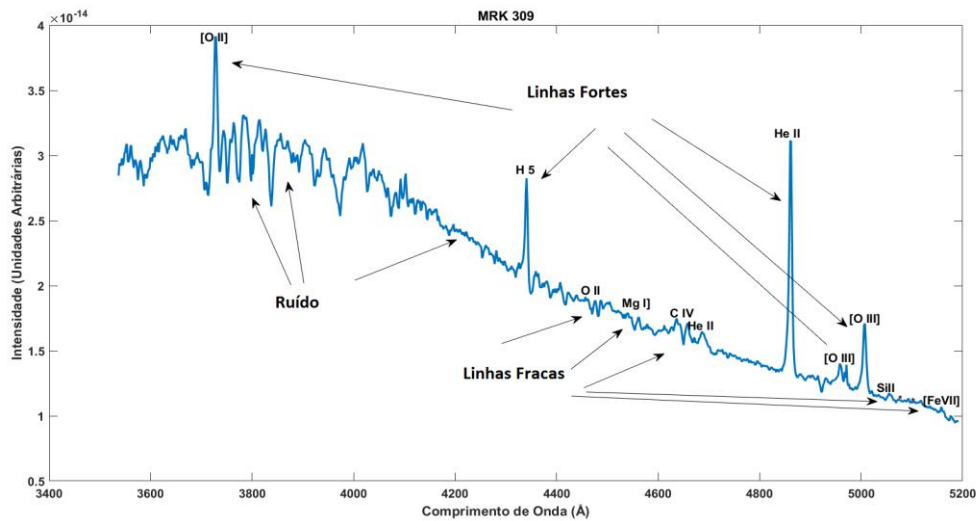


Figura 3.1: Espectro da galáxia MRK 309 com destaque de linhas fortes e fracas.

Logo, podem ser destacados os seguintes problemas:

- ✓ As particularidades dos espectros com níveis de sinal/ruído e resolução variados que dificultam a correta classificação e análise das linhas;
- ✓ O atual volume de dados que torna proibitivo a análise manual dos mesmos;
- ✓ Os atuais programas existentes necessitam de muitos parâmetros para a utilização, dependendo assim da experiência do usuário e demandando uma grande curva de aprendizado, além da necessidade da mudança destes parâmetros para novas análises, o que também consome tempo e pode dar margem a erros.

Tendo em vista, estes problemas, faz-se necessário o desenvolvimento de métodos de análise automática dos espectros, como o desenvolvido nesta pesquisa, para evitar erros ou mesmo auxiliar os especialistas em suas verificações, além de viabilizar o trabalho em grandes bases de dados.

3.1 Delimitações

Inicialmente, para que os dados espectrais possam ser analisados pelos especialistas é necessário realizar, após sua obtenção, um processo inicial de correções conhecido como redução de dados. Nesta etapa, vários processos de correções são realizados para suprimir limitações do próprio instrumento coletor, além dos efeitos atmosféricos, como também o desvio *Doppler* causado pelo movimento da Terra ao

redor do Sol, dentre outros. É importante pontuar que alguns observatórios já possuem *pipelines* para realizar estas correções automaticamente como o Gemini (MAIRE *et al.*, 2010).

Os parâmetros obtidos na observação são diretamente dependentes tanto das características do instrumento do telescópio quanto das condições atmosféricas na data da observação dos referidos dados, o que favorece a utilização desses *pipelines* ou de processos semiautomáticos. Desta forma, este estudo manterá o foco nas condições dos dados posterior aos processos de redução, precisamente na fase inicial que possibilita a identificação das linhas espectrais. Portanto, serão materiais deste trabalho os espectros unidimensionais já calibrados em comprimento de onda e com tais correções já realizadas.

Além disto, as linhas em absorção e as linhas em emissão dos espectros possuem características distintas, linhas em emissão são em geral gaussianas, enquanto linhas em absorção possuem um perfil muitas vezes como combinação de gaussianas e lorentzianas (perfil *Voigt*) (KITCHIN, 1995). As linhas em absorção se formam geralmente em condições de equilíbrio termodinâmico, já as linhas em emissão são formadas em regiões de mecanismos de excitação incomuns (SHARPEE *et al.*, 2003). Assim, geralmente são necessários métodos diferenciados para a identificação automática de ambas.

A partir da correta identificação e medição das linhas em emissão já é possível estimar a temperatura, densidade do gás, abundância química, taxa de formação de estrelas em galáxias, e distinguir galáxias de formação estelar daquelas que possuem um núcleo ativo (STASINSKA, 2007).

Então, devido a importância das linhas em emissão e a necessidade de tratamento diferenciado para a sua identificação, esta pesquisa manterá o foco apenas nestas linhas espectrais. E como cada intervalo do espectro eletromagnético demanda uma tecnologia própria para seu tratamento e obtenção, e ainda, cada objeto astronômico possui linhas com perfis variados, influenciados pelas suas condições físicas intrínsecas, o que exige abordagens próprias, esta pesquisa tratará a princípio com espectros de galáxias na faixa do óptico.

Capítulo 4

Metodologia

“Quanto mais pesquisas fazemos nesse admirável cenário de coisas, mais beleza e harmonia vemos nelas; e mais forte e nítida são as convicções que elas nos dão do ser, do poder e da sabedoria do Arquiteto divino”

-- Steohen Hales

Como visto anteriormente, existe uma crescente necessidade de métodos automáticos para auxílio dos pesquisadores na análise do atual montante de dados astronômicos, sendo a espectroscopia uma das principais áreas de contribuição desta ciência. Por sua vez, esta análise depende inicialmente da correta identificação das linhas espectrais, assim este trabalho teve por objetivo geral o desenvolvimento de uma rotina computacional automática que auxilie o pesquisador a identificar as linhas em emissão em espectros de galáxias na faixa do óptico.

As linhas em emissão nos espectros das galáxias são apresentadas como picos no espectro e têm em geral um perfil gaussiano, ou muito próximo deste (Figura 4.1). De forma distinta, as emissões do ruído, que podem ser gerados por influências diversas como, por exemplo, condições atmosféricas ou por limitações do aparelho coletor, apesar de apresentarem picos não têm como característica o perfil gaussiano. Entretanto, linhas em emissão podem ser tão afetadas pelas condições intrínsecas do meio interestelar, como também pelos processos de obtenção e correção do espectro, que seu perfil se torna severamente comprometido e adicionalmente o ruído também

poderá assumir um perfil gaussiano ou próximo deste, o que inviabiliza uma detecção de linhas em emissão que se baseia apenas no perfil das mesmas. Contudo, esta característica pode ser explorada como um dos descritores de cada pico no espectro para auxiliar na tarefa de classificação.

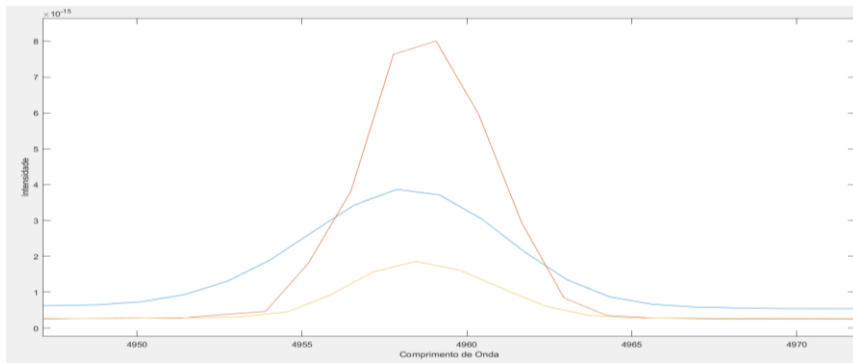


Figura 4.1: Linha em emissão 4959.52[OIII] em três espectros de galáxias distintas.

Como a verificação de apenas uma característica pode não ser viável para esta classificação, se faz necessário observar vários parâmetros para um melhor resultado. Além disso, a correta obtenção desses parâmetros e o desenvolvimento de uma abordagem algorítmica rigorosa baseada em um modelo puramente matemático pode se tornar muito complexo e demorado.

Este tem sido um problema recorrente nas ciências aplicadas, e na atualidade em situações como esta, é frequente o uso de técnicas computacionais de aprendizado de máquina que utilizam dados de exemplo para induzir um modelo de classificação a partir da descoberta de padrões nestes dados (ABDEL-AAL, 2002).

Este método provê uma abordagem vantajosa uma vez que os cálculos intensivos são necessários apenas durante o processo de treinamento do algoritmo para criação do modelo, e não toda vez que uma nova amostra de dados precisar ser processada. Assim, pode ser desenvolvida uma abordagem simples e rápida para a resolução do problema.

Devido a estas vantagens já se faz uso das técnicas de aprendizado de máquina como as RNAs para classificações diversas em dados espectroscópicos (Tu *et al.*, 2008) e em outras áreas, bem como seu uso específico para classificar picos em sinais, como por exemplo, na detecção de radar (WUNSCH, 2015) e em sinais de eletrocardiograma (VIJAYA *et al.*, 1998). O que indica um possível caminho para

uma abordagem específica que utiliza RNA na classificação das linhas em emissão nos espectros de galáxias.

Então, esta abordagem está sendo proposta e avaliada nesta pesquisa, por meio do treinamento supervisionado de Redes Neurais do tipo *Perceptron* Multicamada, e confrontados seus resultados com outros algoritmos de aprendizado de máquina, e com outras abordagens automáticas para o mesmo fim. Para isto, uma sequência de passos precisou ser seguida.

Primeiramente, foi escolhida uma plataforma de desenvolvimento e testes conhecida da comunidade científica e que facilitasse o alcance deste objetivo. Depois, uma base de dados composta por espectros astronômicos de galáxias foi escolhida e processada para permitir o treinamento e avaliação de algoritmos de aprendizado de máquinas que pudessem ser capazes de classificar, como linhas em emissão ou ruído, os picos encontrados nos espectros. Após avaliado, o melhor classificador foi incorporado em uma rotina computacional que atendesse os objetivos desta pesquisa.

4.1 Plataforma de Desenvolvimento

Foi escolhido o MATLAB R2017a como ambiente de desenvolvimento e testes devido a ser uma plataforma amplamente utilizada para análise e manipulação de dados, além da mesma conter ferramentas robustas e de fácil implementação para criação e/ou reutilização de algoritmos de aprendizado de máquina.

4.2 Arquitetura da Rede Neural Artificial

A Rede Neural utilizada nesta pesquisa será do tipo *Perceptron* Multicamada, como função de transferência foi escolhida a tangente sigmoide já que se deseja saídas no intervalo de $[-0,99, 0,99]$. Para medida de avaliação da performance no treino foi escolhido o erro quadrático médio.

A quantidade de neurônios da camada de entrada variou conforme a quantidade de variáveis utilizadas. Para a camada intermediária foram testadas algumas quantidades, mas prevaleceu a quantidade de 10 neurônios, e a camada de saída possui um neurônio que corresponde ao resultado da classificação. Na Figura 4.2

é exemplificada a topologia básica da rede neural.

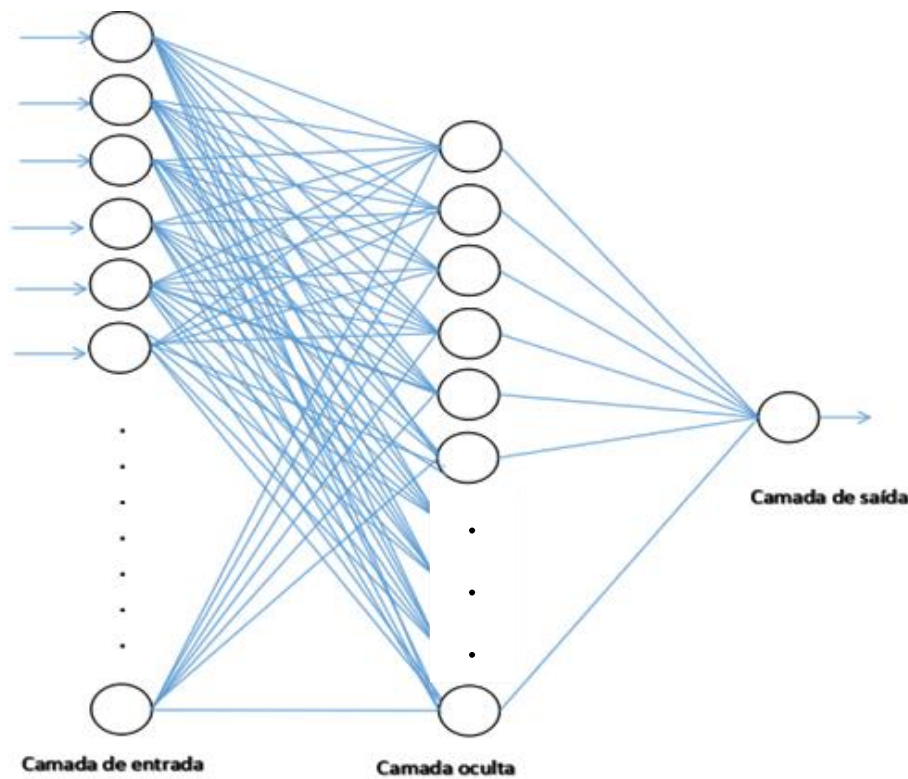


Figura 4.2: Topologia básica da rede neural artificial.

4.3 Base de Dados

A base de dados para esta pesquisa é composta de espectros provenientes de observações de 10 galáxias *Wolf-Rayet* adquiridas no telescópio de 5 m do Monte Palomar cobrindo a faixa de 3600 a 6700 Å do espectro eletromagnético. Com resolução espectral máxima de $\sim 5,6$ Å no azul e $\sim 5,7$ Å no vermelho.

Além destes espectros, também compõem a base de dados os espectros de 2 galáxias que foram obtidos do *New Technology Telescope* (NTT) com 3,6 m, cobrindo a faixa entre 4000 a 6600 Å do espectro eletromagnético e com resolução máxima $\sim 5,9$ Å.

A resolução espectral é a capacidade de separar as linhas espectrais (KITCHIN, 1995), assim, em uma resolução de 5,6 Å é possível ver linhas espectrais que estão separadas no mínimo a 5,6 Å de distância uma da outra, logo, quanto menor for a resolução espectral melhor será a identificação de linhas muito próximas.

As observações em algumas das galáxias que compõem a base foram realizadas

em ângulos diferentes e os espectros para o vermelho e azul foram gerados separados, desta forma a amostra final conteve 26 espectros totalizando 427 classificações de picos como linhas em emissão e 6072 classificações de picos como ruído. Este montante foi dividido em três grupos, Treino, Teste de Avaliação (TA) e Teste Extra (TE), para construção e avaliação do classificador. Um resumo dos dados pode ser visto na Tabela 4.1.

Tabela 4.1: Resumo dos espectros que compõem a base de dados agrupados por galáxias.

Galáxia	Telescópio	Espectros	S/R	Nº Linhas	Grupo
UM 48	Palomar	2	96/ 73	24	Treino
NGC 450	Palomar	2	90/ 77	36	Treino
MRK 712	NTT	1	84	24	TE
NGC 4385	NTT	1	96	23	TE
NGC 4861	Palomar	2	160/ 97	38	Treino
NGC 5430	Palomar	2	140/ 104	29	Treino
NGC 5471	Palomar	2	91/ 75	32	TA
MRK 475	Palomar	2	84/ 70	38	Treino
NGC 6764	Palomar	4	125/ 95	57	Treino/TA
MRK 309	Palomar	2	123/ 90	15	Treino
III Zw 107	Palomar	2	125/ 95	40	Treino
NGC 7714	Palomar	4	99/ 71	71	Treino
Total	-	26	-	427	-

Fonte: Adaptado de Fernandes *et al.*, 2004.

Na Tabela 4.1 a coluna S/R mostra a taxa de sinal ruído para cada espectro, nas galáxias que têm os espectros nas bandas do vermelho e do azul separadas esta taxa é exibida na ordem vermelho/azul respectivamente. A galáxia NGC 6764 foi

observada em dois ângulos de posição 67° e 90° . Os espectros para a observação com 90° foram usados no treino e os espectros para a observação com 67° foram usados no grupo de validação.

Para todos os espectros da base já foi realizado o processo de correção pelo *redshift*, porém as localizações em comprimento de onda previamente conhecidas e tabeladas (FEKLISTOVA *et al.*, 1994) das linhas em emissão não foram utilizadas como informação para alimentar a RNA, pois, em espectros cuja correção do *redshift* não tenha sido realizada, essa informação será desconhecida inicialmente. E deseja-se que a abordagem aqui proposta também possa ser futuramente utilizada no processo de correção do *redshift*.

Em um experimento é necessário reduzir as incertezas, e a escolha apropriada da base de dados pode contribuir para esse fim. Uma base de dados composta por espectros artificiais baseados em simulações é interessante, pois após as devidas validações é possível manipular o seu tamanho facilmente, porém a mesma, além de poder conter erros de *design*, não representa a realidade em todos os aspectos. Por isso, nesta pesquisa deu-se preferência a utilização de uma base de dados proveniente de observações reais e com uma classificação sólida e abalizada das linhas em emissão. A base de dados utilizada nesta pesquisa atende a estes requisitos, pois se trata de espectros reais e sua classificação inicial já foi submetida ao escrutínio científico tanto do autor como da comunidade científica ao ser publicada em Fernandes *et al.* (2004).

4.4 Pré-Processamento

A base de dados passou por uma série de processos para servir como entrada para a RNA, primeiro foi necessário localizar os picos no espectro, e para isto foi utilizado a função *findpeaks* do MATLAB que localiza pontos de máximo local e retorna a localização, a largura a meia altura e a elevação dos picos. Com essas informações e com o uso de um catálogo, baseado nas classificações prévias de um especialista, foi possível identificar as linhas no espectro. Todos os demais picos não classificados como linhas, são classificados como ruído. O pico identificado como linha recebe o valor 0,99 e o ruído -0,99, estes valores estão dentro do intervalo de saída da RNA que utiliza como função de ativação a tangente sigmoide.

O espectro unidimensional é um gráfico de intensidade das linhas em função do seu respectivo comprimento de onda, estes valores são agrupados em um vetor bidimensional, tendo na primeira coluna os valores do comprimento de onda, e na segunda coluna os valores de intensidade. Os picos identificados como linhas, na base de dados utilizada, têm em sua composição entre 5 a 21 amostras de intensidade aproximadamente. Então após alguns testes, tendo como centro a amostra localizada pela função *findpeak* foram escolhidas 21 amostras de intensidade como preditores para a RNA. Esta quantidade foi escolhida após testes feitos com uma variação de 5 a 31 amostras. A Figura 4.3 exibe um pico de linha em emissão.

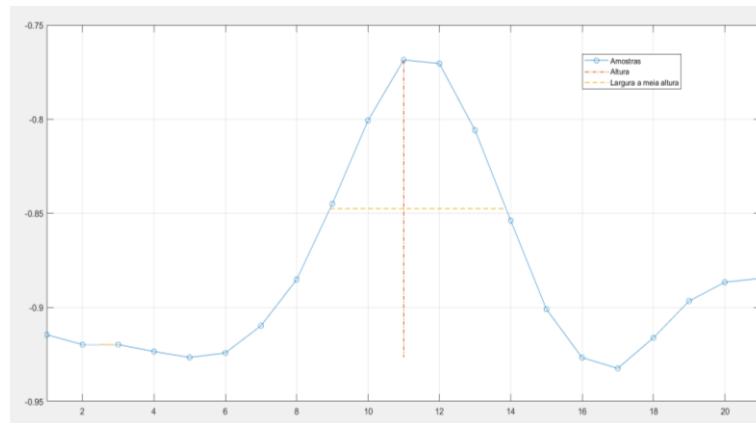


Figura 4.3: Exemplo de um pico de linha em emissão.

Os outros preditores que compõem a entrada da rede são: a mediana e o desvio padrão dos valores de intensidade, a medida da elevação e da largura a meia altura do pico e o coeficiente de correlação entre o pico e uma gaussiana também formada por 21 pontos. A gaussiana é formada pela função *gausswin* do MATLAB, esta função gera uma matriz com a distribuição normal para a quantidade de valores estipulada. O coeficiente de correlação é obtido pela função *corrcoef* do MATLAB, esta função retorna o coeficiente de correlação entre duas matrizes de distribuição normal, logo, quanto mais próximo do perfil gaussiano o pico do espectro estiver, maior será o valor do coeficiente.

As medidas dos preditores precisam ser uniformizados, desta forma algumas normalizações foram feitas. Os valores de intensidade foram normalizados no intervalo de -1 a 1 (Equação 1). E os valores da largura e da elevação foram normalizados entre 0 a 1 (Equação 2). O desvio padrão e mediana são obtidos dos

valores normalizados e o coeficiente de correlação é utilizado sem normalização por possuir uma variação de -1 a 1. As normalizações são feitas por espectro e não por toda a base. Testes prévios apontaram que a Rede Neural obtém melhor resultado de classificação após a normalização da entrada.

(1)

$$y_0(n) = 2 * \frac{y(n) - \min(y(n))}{(\max(y(n)) - \min(y(n))) - 1}$$

(2)

$$y_0(n) = \frac{y(n) - \min(y(n))}{\max(y(n)) - \min(y(n))}$$

Após todo este processo, a base de dados está pronta para servir de entrada para a RNA.

4.5 Método e Medidas de Avaliação

Após o pré-processamento, os dados foram divididos em três conjuntos principais: Treino, Teste de avaliação e Teste extra. O resumo dos dados de cada grupo é apresentado na Tabela 4.2.

Tabela 4.2: Resumo dos dados de treino e testes.

Grupo	Espectros	Linhas	Ruído	Linhas Fortes	Linhas Fracas
Treino	20 76,9%	321 69,6%	4351 72,0%	22 29,7%	298 81,8%
Teste de avaliação	4 15,4%	59 12,8%	866 14,0%	17 23%	42 11,5%
Teste extra	2 7,7%	47 17,6%	855 14,0%	23 47,3%	24 6,7%
Total	26 100%	427 100%	6072 100%	74 100%	364 100%

Como existe um desbalanceamento entre a quantidade de linhas em emissão e de ruído em cada espectro, foi necessário superamostrar as linhas para balancear sua quantidade e assim possibilitar uma melhor generalização do classificador. Desta forma a quantidade de linhas em emissão da amostra de treino passou de 321 para

4379.

Para cada grupo foram consideradas como linhas fortes as linhas que sobressaem em altura ao ruído, e linhas fracas são as linhas com altura igual ou inferior a altura máxima do ruído conforme exemplificado na Figura 4.4.

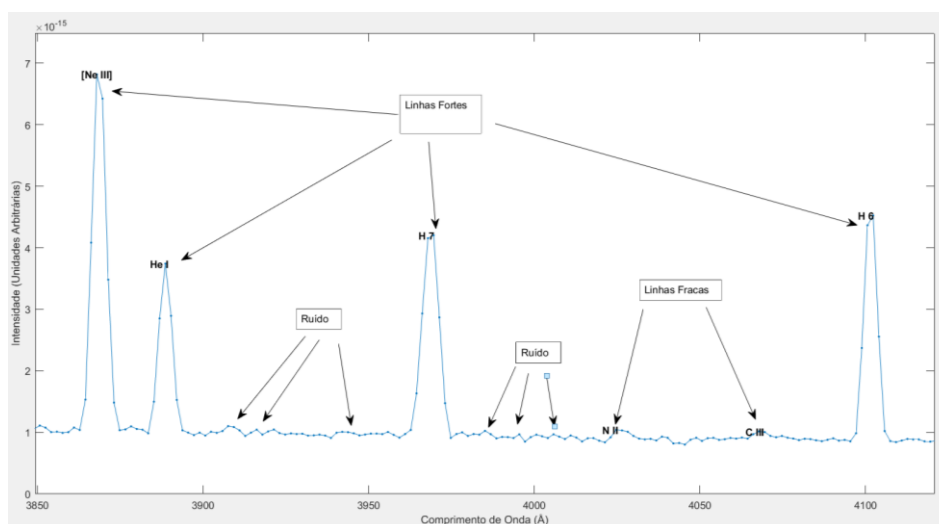


Figura 4.4: Exemplo de linhas fortes, fracas e ruído.

Com a finalidade de se obter uma estimativa da capacidade de classificação da abordagem testada ao longo do conjunto de dados, foi utilizado o *cross-validation* que é uma das abordagens mais utilizadas para este tipo de estimativa. Neste processo, o conjunto de dados é dividido em k subconjuntos de tamanhos aproximadamente iguais (k *folds*), e então o modelo proposto é treinado k vezes de forma independente. Em cada iteração são utilizados $k-1$ subconjuntos como dados de treino e a amostra restante é utilizada como teste para avaliação do treino. Assim é possível obter uma estimativa média de avaliação do classificador em todo o conjunto (JAPKOWICZ, 2011), e espera-se que este seja o comportamento do classificador em amostras fora do conjunto de treino.

Então o conjunto de treino é subdividido a cada *fold* do *cross-validation*, para cada *fold* subconjuntos aleatórios são escolhidos para treino (e validação), que corresponde a 90% da amostra de treino e os 10% restantes da amostra são usados no teste. Esta amostra de testes usada no *cross-validation* a cada *fold* será chamada no restante do texto de amostra de teste do treino (T1).

O conjunto de teste de avaliação serve como um teste controlado para

avaliação do classificador, uma vez que ele é composto pelos espectros completos de duas galáxias diferentes que foram escolhidas de forma aleatória da base de dados. Desta forma é possível avaliar a capacidade de generalização da RNA em uma amostra completa. Esta amostra será chamada de amostra de teste de avaliação (T2) no decorrer do texto e nesta não foi realizada a superamostragem das linhas em emissão.

Além dos testes realizados pela validação cruzada, foram separados dados de espectros que possuem algumas características distintas das existentes na amostra de treino e teste, como por exemplo a resolução espectral e taxa de sinal/ruído. Assim proporciona-se uma avaliação de como a abordagem funciona para espectros totalmente desconhecidos do treino, obtidos por instrumentos diferentes e com resolução espectral inferior. Esta amostra será nomeada de amostra de teste extra (T3) no decorrer do texto.

Após definidos e separados cada conjunto de dados, é necessário definir a melhor estratégia para treinar a RNA. O método de treino escolhido afeta diretamente a capacidade de classificação desta, e não existe um método definido como melhor para cada tipo de aplicação de uma RNA. Em geral se escolhe este método de forma experimental, por isso, nesta pesquisa alguns experimentos foram realizados com o propósito de avaliar um método de treinamento que gere uma RNA capaz de classificar corretamente não apenas os dados de treino, mas principalmente dados não apresentados durante este processo.

O processo de treino foi avaliado por meio de *cross-validation* com 10 *folds* e AUC verificando as médias aritméticas (\bar{x}) e desvio padrão (S) das medidas de: *Recall*, *Precisão*, *Acurácia* e *Area Under the Curve* (AUC). Essas medidas foram tomadas para a classificação obtida no percentual de teste T1 reservado da amostra de treinamento, sobre a amostra para testes de avaliação T2, bem como na amostra de teste extra T3.

Estas medidas foram escolhidas, pois neste trabalho é realizada uma classificação binária, ou seja, existe a classe positiva que são os picos que representam linhas em emissão, e existe a classe negativa que são os picos que representam o ruído. Quando se classifica corretamente a classe positiva se tem um verdadeiro

positivo (*True Positive* - TP). Quando se classifica um pico de ruído como linha se tem um falso positivo (*False Positive* - FP). A classificação correta de um ruído é chamada de verdadeiro negativo (*True Negative* - TN), já classificar uma linha como ruído gera um falso negativo (*False Negative* - FN). As medidas escolhidas possibilitam avaliar o classificador em diversos ângulos.

Neste sentido, o *Recall*, também conhecido como sensibilidade, representa o percentual de amostras positivas classificadas corretamente em relação às amostras positivas reais, conforme a Equação (3).

(3)

$$Recall = \frac{TP}{TP + FN}$$

A Precisão se refere ao percentual de amostras positivas corretamente classificadas sobre o total de amostras classificadas como positivas conforme a Equação (4).

(4)

$$Precisão = \frac{TP}{TP + FP}$$

A Acurácia é calculada somando-se os verdadeiros positivos com os verdadeiros negativos e dividindo-se o resultado desta soma pela soma do total de amostras das classes positivas (P) com o total de amostras das classes negativas (N) conforme Equação (5).

(5)

$$Acurácia = \frac{TP + TN}{P + N}$$

A AUC permite quantificar a exatidão do classificador, pois ela representa o desempenho deste classificador em todas as proporções de custo possíveis (JAPKOWICZ, 2011). Esta medida é calculada para a área existente sob a Curva de Característica de Operação do Receptor (Curva ROC), este tipo de curva foi desenvolvido no campo da comunicação como uma forma de demonstrar a relação entre sinal/ruído. A Curva ROC pode ser visualizada na forma de um gráfico de sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos

(MARGOTTO, 2009) como exibido na Figura 4.5.

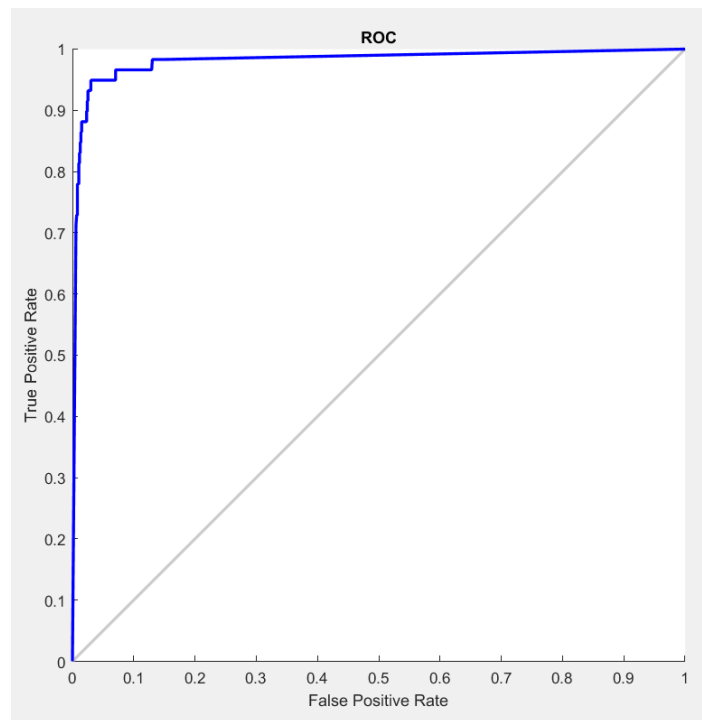


Figura 4.5: Exemplo de uma Curva ROC.

A linha diagonal pontilhada corresponde a uma classe que é positiva ou negativa, aleatoriamente. A Curva ROC auxilia na escolha do limiar de decisão para um classificador, ou seja, o melhor valor limite para o ponto de corte do classificador. Pois, ela evidencia os valores para os quais existe maior otimização da sensibilidade em função da especificidade que corresponde ao ponto em que se encontra mais próxima do canto superior esquerdo do diagrama, uma vez que o índice de verdadeiros positivos é 1 e o de falsos positivos é zero (MARGOTTO, 2009). Por sua vez, quanto maior for a área sob a curva (AUC), melhor será considerado o classificador.

Após escolhidas as métricas de avaliação, é necessário definir e avaliar os parâmetros para o treino da RNA, pois a avaliação de como cada parâmetro afeta o treinamento permite a escolha adequada, bem como a avaliação da interação entre cada conjunto de parâmetros. Porém, analisar a variação dos parâmetros em todas as combinações possíveis para os experimentos desejados nesta pesquisa alcançaria milhares de experimentos, o que torna essa possibilidade muito custosa no quesito tempo e ainda dificulta a análise, então os experimentos foram organizados em blocos.

Nestas as interações de alguns parâmetros são observadas e avaliadas, desta forma o grupo de parâmetros que obteve melhor avaliação passa a ser utilizado no próximo bloco.

Para que um determinado conjunto de parâmetros em um bloco de experimentos possa ser considerado melhor em relação ao outro, a média e desvio padrão para cada medida sobre a amostra de testes T1 foi observado, como também foi realizado o teste estatístico não paramétrico de *Friedman*, pois a amostra é pequena e pode assumir uma distribuição diferente da Normal. Complementando o teste de *Friedman*, foi aplicada uma análise *post hoc* pelo método de comparação de *Bonferroni*, sendo assim possível verificar em uma análise com (n) domínios e (k) classificadores se existem diferenças estatisticamente significativas dos parâmetros/classificadores testados.

Em situações em que não existiam diferenças estatisticamente significantes para todas as medidas, foi observado como medida principal o balanceamento entre o *recall* e a precisão juntamente com as características dos parâmetros em teste.

4.6 Experimentos

O desempenho na classificação desejada obtido por uma RNA varia conforme o processo de treinamento utilizado, pois a RNA é sensível aos parâmetros de treinamento e aos preditores utilizados. Vários experimentos foram realizados para analisar a influência dos mesmos nas medidas de avaliação da rede e assim foi possível escolher os que apresentaram os melhores resultados.

Os parâmetros submetidos à análise foram os que geralmente não apresentam um valor teórico ideal na literatura e que na maioria das vezes são escolhidos por observações empíricas, são eles: algoritmo de treino (2 alternativas), quantidade de neurônios na camada intermediária (5, 10, 20), quantidade de épocas de treino (100, 500, 1000), retrainar a rede com/sem reinicialização dos pesos no processo de *cross-validation* e variação de preditores a serem utilizados. Estes parâmetros foram analisados em três blocos.

No primeiro bloco de experimentos foi verificado o uso de dois algoritmos de

treinamento para RNA: *Bayesian Regularization Backpropagation* (BRB) e *Levenberg-Marquardt Backpropagation*(LMB), ambos disponíveis no MATLAB. Existem vários algoritmos que visam melhorar a aprendizagem de uma RNA em seu treinamento, e em geral é desejável que durante o treinamento a RNA não se limite a soluções em mínimos locais, buscando se aproximar do mínimo global, além de não se superadaptar aos dados de treino para que assim venha a ter uma boa generalização. A generalização é a capacidade da rede fazer boas classificações em dados não apresentados no treinamento (DEMUTH *et al.*, 2014). Desta forma foram escolhidos para avaliação dois algoritmos de treinamento que se propõem a alcançar estes requisitos.

O algoritmo *Levenberg-Marquardt* é um método aprimorado de otimização por *backpropagation*. Na versão básica do *backpropagation* o aprendizado da rede é realizado por meio de processos iterativos de gradiente descendente em que os ajustes nos pesos são feitos por retro propagação a partir de um sinal de erro calculado pela comparação entre o valor de saída alcançado e o desejado (Ooyen e Nienhuis, 1992). Já o algoritmo LMB usa como método de otimização o método de *Levenberg-Marquardt* que introduz um parâmetro para estabilizar o treinamento. Esta estabilização é feita por meio do ajuste da aproximação, ou seja, incrementando o parâmetro caso a função de erro, que se deseja minimizar, seja maximizada ou decrementando-o caso a mesma seja diminuída. Assim o algoritmo busca um equilíbrio entre velocidade e convergência (DEMUTH *et al.*, 2014).

O outro algoritmo testado, BRB, além das características citadas para o LMB, utiliza a Regularização *Bayesiana* para minimizar a combinação dos erros quadráticos e os pesos, visando produzir uma rede generalizada e com respostas mais suaves, para isto são introduzidos termos estabilizadores, que são: o termo de erro padrão, para medir o erro padrão entre a resposta desejada e a obtida, e o termo de regulamentação que penaliza a complexidade (LIVINGSTONE, 2008).

Em adição à verificação do algoritmo de treinamento, neste bloco ainda será avaliada a quantidade de neurônios para a camada intermediária da RNA. Este parâmetro é importante, pois em geral quanto mais neurônios uma rede possuir, maior é a possibilidade de haver um sobre ajuste no treino, além de elevar a

complexidade da rede, e quanto menor esta quantidade, poderá haver um subajuste da mesma (DA SILVA *et al.*, 2016). Em geral, a quantidade ideal de neurônios na camada intermediária varia de problema para problema e é determinada de maneira empírica.

No segundo bloco é verificado quantas épocas de treinamento devem ser utilizadas e se os pesos e *bias* devem ser reinicializados em cada *fold*. A quantidade de épocas de treinamento é um dos fatores utilizados para determinar o fim do treinamento de uma RNA. A parada muito cedo pode favorecer um subajuste, ou seja, o treinamento termina antes da Rede Neural alcançar uma boa taxa de aprendizado, já a parada tardia pode favorecer o superajuste. Então um ponto de equilíbrio deve ser encontrado (DA SILVA *et al.*, 2016).

Com relação à escolha do momento de parada, outro fator que pode ser utilizado é a quantidade de épocas de checagem de validação, ou seja, um subconjunto é escolhido dos dados de treino para que a cada iteração seja verificado o aprendizado da rede para estes dados de validação. Quando o erro aumenta sucessivamente em várias iterações o treino é encerrado e os pesos da época em que houve o menor erro de validação é utilizado como resultado do treino (DEMUTH *et al.*, 2014).

A Figura 4.6 ilustra um exemplo da influência da parada utilizando a checagem de validação. Na parte inferior é exibido o progresso do desempenho do treinamento/checagens de validação. No ponto (a) é observado que a taxa de erro no aprendizado continuaria caindo no treinamento, mas que a partir deste ponto, esta taxa aumentaria no conjunto de validação (ponto b). O gráfico superior esquerdo revela que neste ponto a rede tem um bom ajuste para a função desejada, mas caso este treinamento continuasse o ajuste se deterioraria conforme visto no gráfico superior direito. Os valores testados para a quantidade de épocas de treino\checagem de validação foram: 100/20, 500/100, 1000/200 e 1000/500.

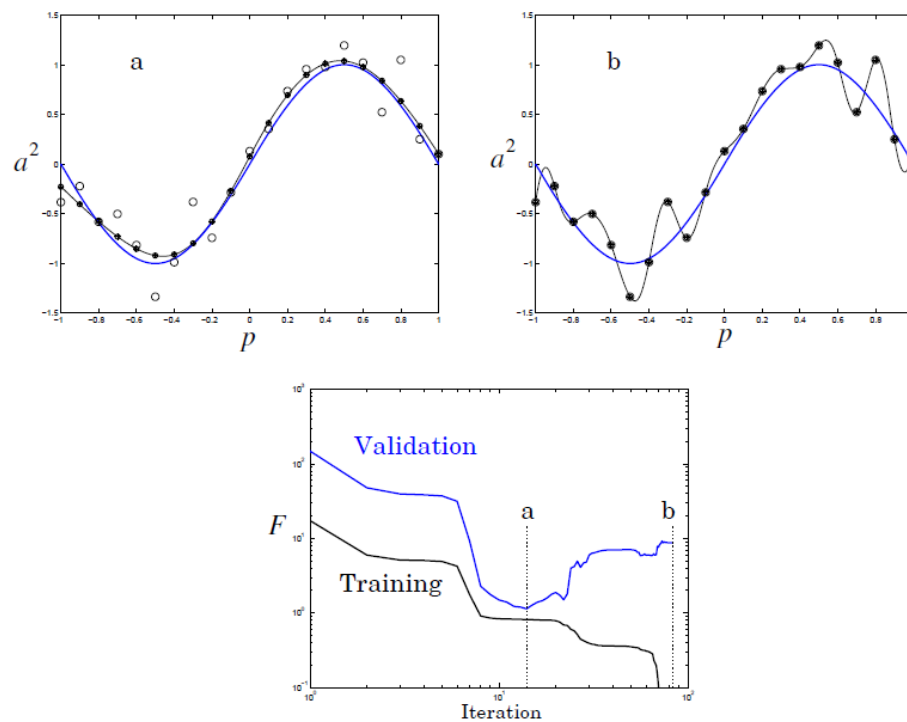


Figura 4.6: Processo de parada por checagens de validação. Fonte: (DEMUTH *et al.*, 2014).

Ainda neste bloco foi verificado como a reinicialização dos pesos e *bias* a cada *fold* influencia no aprendizado da RNA, pois a depender do ponto de inicialização a rede pode se dirigir, no espaço de soluções, para um ponto de solução ou outro. A Figura 4.7 ilustra no espaço de soluções a variedade de soluções existentes, e é possível ver que existe um mínimo global, que representa a melhor solução possível a ser encontrada, e também existem outras soluções possíveis (mínimos locais). Quanto mais a rede converge para próximo do mínimo global, melhor. E esta convergência depende de onde a matriz de pesos ($W^{(n)}$) foi inicializada (DA SILVA *et al.*, 2016).

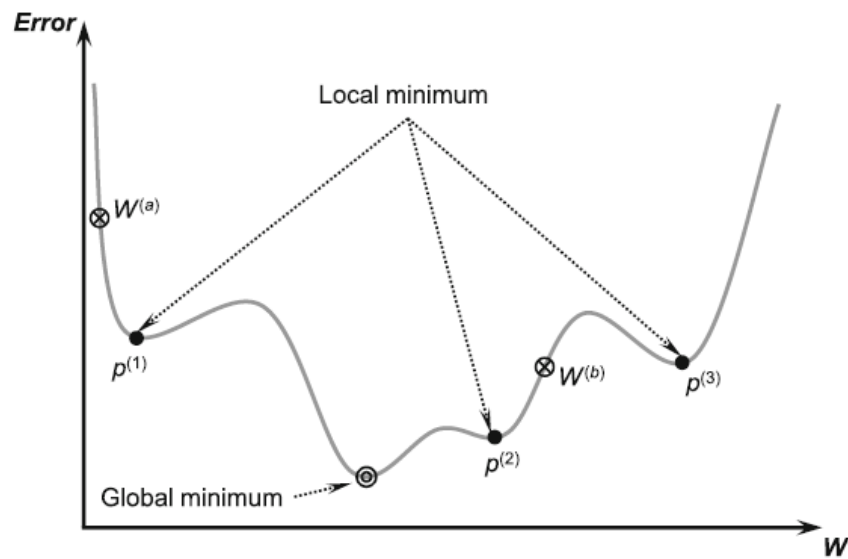


Figura 4.7: Variedade de soluções no espaço de soluções. Fonte: (DA SILVA *et al.*, 2016).

A depender dos métodos de treinamento, a rede pode se deslocar de um ponto de mínimo local para outro que oferece uma melhor solução. Quando a RNA é iniciada a cada *fold*, seu local inicial pode ser aleatório, mas quando os valores dos pesos e *bias* são preservados de um *fold* a outro, a rede utiliza estes valores como sugestão inicial para o novo espaço de soluções da *fold* em questão e pode se deslocar para um outro local mínimo, ou de preferência para o mínimo global.

Apesar da simples continuação do treino ou reinicialização do mesmo, não há garantias que o outro local seja melhor que o primeiro. Porém, como a cada *fold* o desempenho daquele ponto do treinamento é armazenado, é possível saber se esta estratégia melhorou ou não o aprendizado.

No terceiro e último bloco de experimentos é avaliado como os preditores escolhidos impactam o aprendizado da Rede Neural. Para isto os preditores foram divididos em três grupos e o treinamento foi avaliado observando as respostas da rede tendo como entrada as variações desses grupos. O primeiro grupo (G1) diz respeito a quantidade de amostras de intensidade do pico que forma a linha em emissão, o segundo (G2) possui medidas relacionadas a essas amostras, que são a mediana e o desvio padrão dos valores de intensidade e o terceiro grupo (G3) é composto por medidas que são relacionadas ao formato do pico como: elevação, largura e sua correlação com uma gaussiana.

Após a avaliação e escolha dos parâmetros para o melhor procedimento de treinamento, a RNA foi treinada com toda a amostra de treino para garantir um aprendizado sobre um maior número de exemplos. E em seguida foram feitas as verificações das medidas para os testes T2 e T3.

Com o teste de avaliação e o teste extra concluídos, os resultados obtidos foram confrontados com os resultados de outros classificadores. São eles o *Support Vector Machine* (SVM) e Regressão Logística.

O SVM é um algoritmo de aprendizado de máquina que localiza uma fronteira de separação (hiperplano) entre duas classes, visando maximizar a margem de separação que se encontra entre os pontos mais próximos da superfície de decisão, conhecidos como vetores de suporte (CRISTIANINI *et al.*, 2000). A função de *kernel* utilizada para o SVM foi a linear, por obter melhores resultados em um teste prévio nesta pesquisa, do que a gaussiana ou a polinomial.

A Regressão Logística foi implementada com a seleção por etapas, isto quer dizer que o modelo foi construído de forma incremental, em que a cada incremento os preditores eram incluídos ou removidos. Isto permite verificar quais preditores são mais relevantes estatisticamente para o modelo, o que pode ajudar a obter um modelo com uma melhor generalização (DREISEITL *et al.*, 2002).

Além da comparação com os classificadores citados, uma comparação foi realizada entre a rotina computacional desenvolvida nesta pesquisa e com o ALFA (WESSON, 2016), que foi a rotina computacional, encontrada na literatura, mais atual e com a proposta mais semelhante à desta pesquisa.

Capítulo 5

Resultados e Discussões

“Os argumentos mais fortes não provam nada, desde que as conclusões não são verificadas pela experiência. Ciência experimental é a rainha das ciências e da meta de todas as especulações.”

-- Roger Bacon

Nesta seção estão sendo apresentados os resultados para cada bloco de experimentos realizados e são feitas as considerações sobre os mesmos.

5.1 Treinamento e Testes de Validação

No primeiro bloco de experimentos foram avaliados dois algoritmos de treinamento (AT) para redes neurais: O *Bayesian regularization backpropagation* (BRB) e o *Levenberg-Marquardt backpropagation* (LMB). Além disso, para cada algoritmo foi observado sua interação com redes treinadas com 1000 épocas de treino e 500 épocas de validação e com 5, 10 e 20 neurônios na camada intermediária representado com a sigla Q.N na Tabela 5.1. Nesta tabela é possível ver as médias e desvio padrão para as medidas alcançadas em cada *fold* para a amostra de teste no treino que recebeu a sigla T1.

Tabela 5.1: Resultados do treinamento com os parâmetros do bloco 1 de experimentos.

AT	Q.N	<i>Recall</i>	<i>Precisão</i>	<i>Acurácia</i>	<i>AUC</i>
		$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$
(1) BRB	5	0.981±0.011	0.958±0.008	0.969±0.007	0.979±0.004
(2) LMB	5	0.984±0.006	0.953±0.007	0.968±0.003	0.972±0.004
(3) BRB	10	0.987±0.004	0.963±0.012	0.974±0.006	0.980±0.003
(4) LMB	10	0.986±0.007	0.965±0.006	0.975±0.004	0.980±0.004
(5) BRB	20	0.987±0.006	0.969±0.007	0.977±0.002	0.983±0.003
(6) LMB	20	0.988±0.005	0.962±0.007	0.974±0.005	0.980±0.006

Analisando a Tabela 5.1 é possível verificar que as diferenças entre as medidas para todos os experimentos foram muito próximas, então uma análise estatística é necessária para verificar se existem diferenças significativas entre eles.

Primeiro foi realizado o teste estatístico de *Friedman* para o *recall*, e como visto na Tabela 5.2 o valor de $\chi^2 = 6,25$ é menor do que o valor crítico (11,07) para rejeitar a hipótese nula (H_0) para 0,5 de significância em um teste realizado em 10 domínios ($n=10$) e 6 algoritmos ($k=6$) conforme valores tabelados disponíveis em Japkowicz (2011). Adicionalmente pode ser observado que o *p-value* foi alto o que também indica que não é possível rejeitar H_0 , ou seja, em todos os experimentos não houve diferença estatisticamente significativa para o *recall*.

Tabela 5.2: Valores estatísticos do teste de *Friedman* para o *recall* do bloco 1 de experimentos.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	<i>p-value</i>
Teste	6,25	0,2826
H_0 para 0,5	11,07	

A Tabela 5.3 exhibe o teste estatístico para a medida de precisão, e conforme exibido, o valor de $\chi^2 = 22,21$ é maior do que o valor crítico (11,07) para rejeitar a

hipótese nula (H_0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno é possível rejeitar H_0 , ou seja, existem diferenças estatisticamente significativa nos experimentos para as medidas de precisão.

Tabela 5.3: Valores estatísticos do teste de *Friedman* para a *precisão* do bloco 1 de experimentos.

Nº de Algoritmos=6	χ^2	p-value
Nº de domínios=10		
Teste	22,21	0,0004
H_0 para 0,5	11,07	

Agora, é necessário realizar uma análise *post hoc* pelo método de *Bonferroni* para verificar quais parâmetros nos experimentos contribuíram para essa variação. A Tabela 5.4 exibe as comparações múltiplas entre os experimentos para esta verificação e na Figura 5.1 é possível visualizar essas variações.

Tabela 5.4: Valores estatísticos do método de *Bonferroni* para a precisão do bloco 1 de experimentos.

Experimentos		L. Inferior	Média	L. Superior	<i>P-Value</i>
1	2	-1,699	0,75	3,199	1,000
1	3	-3,849	-1,4	1,049	1,000
1	4	-4,699	-2,25	0,199	0,105
1	5	-4,849	-2,4	0,049	0,060
1	6	-3,449	-1	1,449	1,000
2	3	-4,599	-2,15	0,299	0,149
2	4	-5,449	-3	-0,551	0,005
2	5	-5,599	-3,15	-0,701	0,002
2	6	-4,199	-1,75	0,699	0,539
3	4	-3,299	-0,85	1,599	1,000
3	5	-3,449	-1	1,449	1,000
3	6	-2,049	0,4	2,849	1,000
4	5	-2,599	-0,15	2,299	1,000
4	6	-1,199	1,25	3,699	1,000
5	6	-1,049	1,4	3,849	1,000

Na Tabela 5.4 as duas primeiras colunas representam os experimentos que são testados por pares, a terceira e quinta colunas representam os limites inferior e superior para o intervalo de 95% de confiança para a diferença média verdadeira, a quarta coluna mostra a diferença entre a média dos grupos estimados e a sexta coluna exibe o *p-value* do teste de hipótese para a diferença média ser igual a zero. Desta forma, quanto mais próximo de zero é este valor, maior é a significância da diferença entre os experimentos. A diferença entre os grupos também pode ser vista na Figura 5.1.

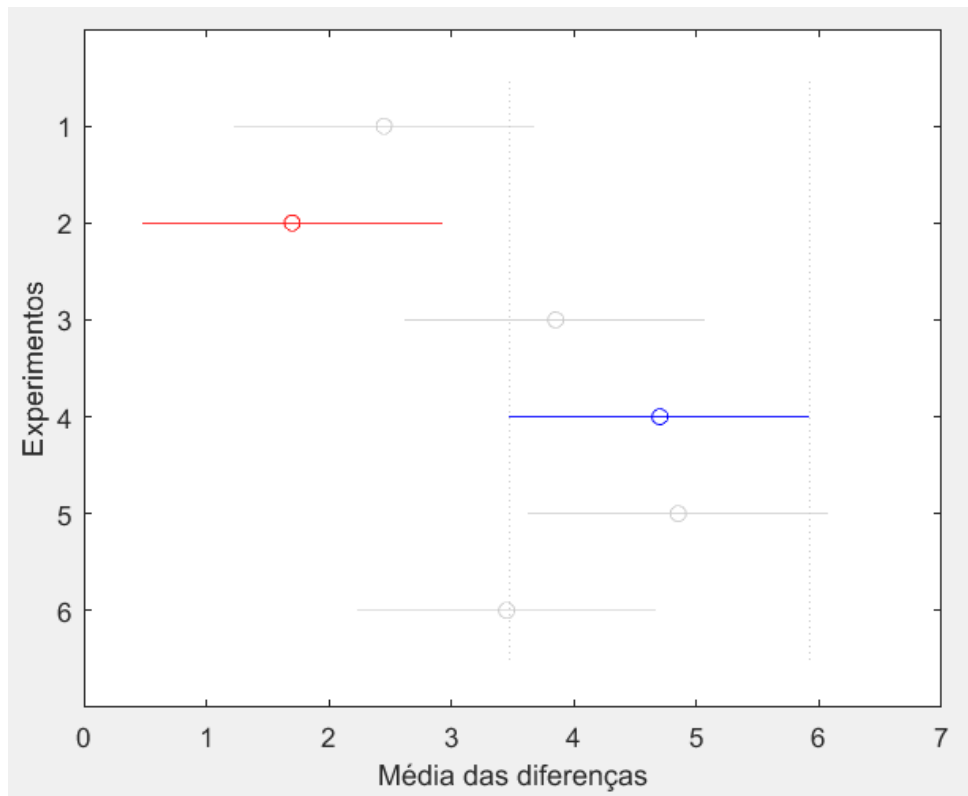


Figura 5.1: Gráfico para o teste de *Bonferroni* para a precisão do bloco 1 de experimentos.

Na Figura 5.1 e na Tabela 5.4 é visto que os experimentos 1, 3, 4, 5 e 6 não possuem diferença significativa, porém o experimento 2 é diferente e contribui menos para a melhoria da precisão, os experimentos 4 e 5 são os mais semelhantes, sendo então os seus parâmetros os que alcançaram uma melhor precisão neste bloco.

A próxima medida a ser verificada é a acurácia, e na Tabela 5.5 é visto as estatísticas do teste. O valor de $\chi^2 = 21,23$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H_0) para 0,5 de significância. E observando o *p-value* que foi bem pequeno (0,0007) é possível rejeitar H_0 , ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas de acurácia.

Tabela 5.5: Valores estatísticos do teste de *Friedman* para a acurácia do bloco 1 de experimentos.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	<i>p-value</i>
Teste	21,23	0,0007
H_0 para 0,5	11,07	

Visto que existem diferenças significativas uma análise *post hoc* pelo método de *Bonferroni* foi executado para verificar quais parâmetros nos experimentos contribuíram para essa variação. A Tabela 5.6 exhibe as comparações múltiplas entre os experimentos para esta análise e na Figura 5.2 é possível visualizar suas variações.

Tabela 5.6: Valores do método de *Bonferroni* para a acurácia do bloco 1 de experimentos.

Experimentos	L. Inferior	Média	L. Superior	<i>P-Value</i>
1 2	-1,649	0,75	3,149	1,000
1 3	-3,999	-1,6	0,799	0,754
1 4	-4,099	-1,7	0,699	0,563
1 5	-4,849	-2,45	-0,051	0,041
1 6	-3,699	-1,3	1,099	1,000
2 3	-4,749	-2,35	0,049	0,061
2 4	-4,849	-2,45	-0,051	0,041
2 5	-5,599	-3,2	-0,801	0,001
2 6	-4,449	-2,05	0,349	0,182
3 4	-2,499	-0,1	2,299	1,000
3 5	-3,249	-0,85	1,549	1,000
3 6	-2,099	0,3	2,699	1,000
4 5	-3,149	-0,75	1,649	1,000
4 6	-1,999	0,4	2,799	1,000
5 6	-1,249	1,15	3,549	1,000

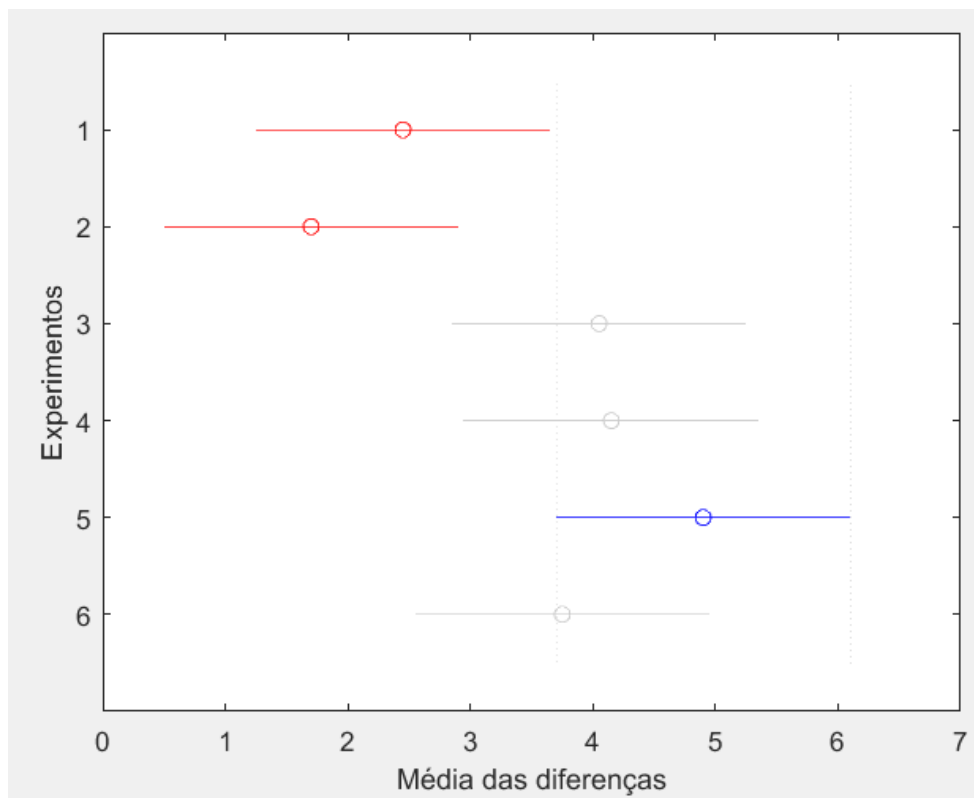


Figura 5.2: Gráfico do método de *Bonferroni* para a acurácia do bloco 1 de experimentos.

De acordo ao teste de *Friedman* e a análise pelo método de *Bonferroni* os efeitos dos experimentos 3, 4, 5 e 6 foram estatisticamente semelhantes, porém diferentes dos experimentos 1 e 2, sendo estes últimos os que menos contribuíram para o aumento da acurácia, e os parâmetros dos experimentos 4 e 5 novamente foram os que mais efeitos positivos tiveram sobre a acurácia.

Como última verificação do primeiro bloco de experimentos foi aplicado inicialmente o teste estatístico para avaliar se existem diferenças estatisticamente significantes para os efeitos dos parâmetros escolhidos sobre a AUC. A Tabela 5.7 exhibe as estatísticas do teste de *Friedman*.

Tabela 5.7: Valores estatísticos do teste de *Friedman* para a AUC do bloco 1 de experimentos.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	p-value
Teste	20,69	0,0009
H0 para 0,5	11,07	

Na Tabela 5.7 mostra-se que o valor de $\chi^2 = 20,69$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E observando o *p-value* que foi bem pequeno (0,0009) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas da AUC. Então, visto que existem diferenças significativas, a análise *post hoc* pelo método de *Bonferroni* foi realizada para verificar quais parâmetros nos experimentos contribuíram para essa variação. A Tabela 5.8 exhibe as comparações múltiplas entre os experimentos para esta análise e na Figura 5.3 é possível visualizar essas variações.

Tabela 5.8: Valores do método de *Bonferroni* para a AUC do bloco 1 de experimentos.

Experimentos	L. Inferior	Média	L. Superior	<i>P-Value</i>
1 2	-0,156	2,3	4,756	0,090
1 3	-2,556	-0,1	2,356	1,000
1 4	-2,056	0,4	2,856	1,000
1 5	-3,856	-1,4	1,056	1,000
1 6	-1,856	0,6	3,056	1,000
2 3	-4,856	-2,4	0,056	0,062
2 4	-4,356	-1,9	0,556	0,347
2 5	-6,156	-3,7	-1,244	0,0001
2 6	-4,156	-1,7	0,756	0,632
3 4	-1,956	0,5	2,956	1,000
3 5	-3,756	-1,3	1,156	1,000
3 6	-1,756	0,7	3,156	1,000
4 5	-4,256	-1,8	0,656	0,472
4 6	-2,256	0,2	2,656	1,000
5 6	-0,456	2	4,456	0,252

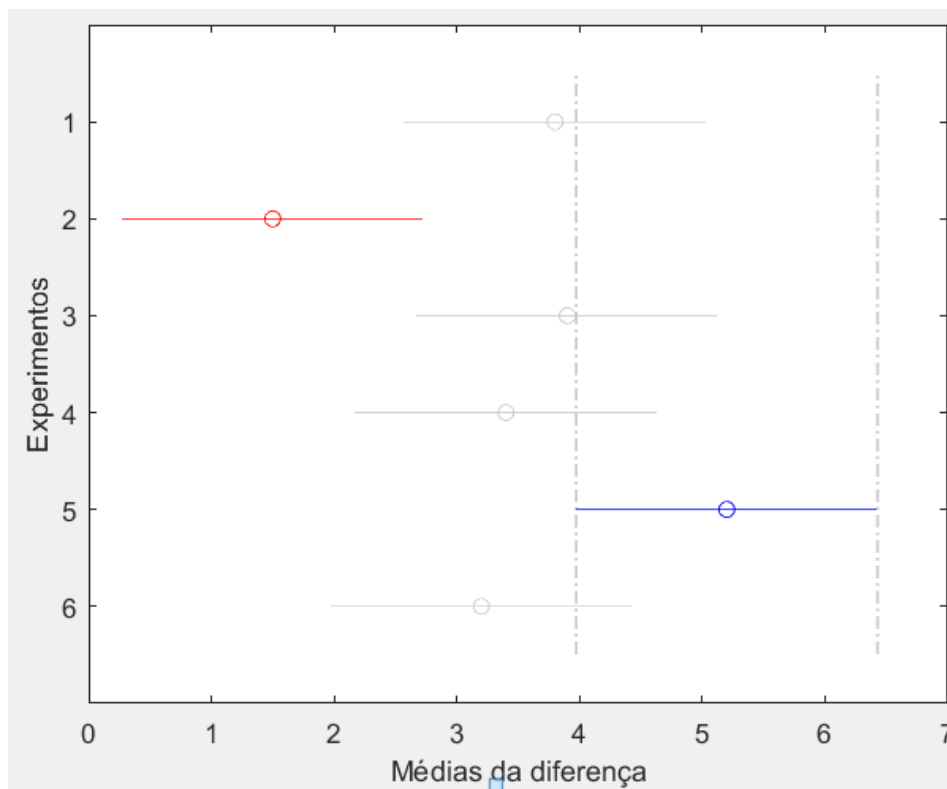


Figura 5.3: Gráfico do método de *Bonferroni* para a AUC do bloco 1 de experimentos.

Nesta análise foi observado que os experimentos 1, 3, 4, 5, e 6 são estatisticamente similares em seus efeitos sobre a acurácia, enquanto que o experimento 2 tem variações em relação ao 5, contribuindo menos para a melhoria desta medida.

Após os testes estatísticos serem executados e suas respectivas análises terem sido realizadas, é possível dizer que os experimentos com 5 neurônios para os dois algoritmos testados tiveram no geral as medidas mais baixas (Tabela 5.1), os demais foram considerados estatisticamente semelhantes em seus efeitos, porém, como o treino com o algoritmo LMB com 10 neurônios converge mais rápido e é menos provável de se superajustar aos dados de treino do que com 20 neurônios, estes serão os parâmetros que continuarão a ser utilizados no próximo bloco de experimentos.

No segundo bloco foi avaliado o treinamento da rede com parâmetros de épocas de treino diferentes (ET) e com (identificado com um “x” na Tabela 5.9) ou sem reinicialização (R) dos pesos e *bias* a cada *fold*, ou seja, cada *fold* começará o treinamento com os pesos usados na *fold* anterior.

Anteriormente para cada quantidade de épocas de treino foi escolhida 50% desta quantidade para épocas de checagem de validação, neste bloco foram testados os efeitos de 20% das épocas de treino para essa checagem. Assim, para o treino com 100 épocas, foram utilizadas 20 épocas de checagem de validação e assim por diante. A checagem de validação serve como um parâmetro de controle para evitar a superadaptação da rede aos dados de treino. A Tabela 5.9 exibe as medidas de avaliação para os experimentos do bloco 2.

Tabela 5.9: Medidas de avaliação para os experimentos do bloco 2.

ET	R	Teste	<i>Recall</i>	Precisão	Acurácia	AUC
			$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$
100/20	x	T1	0.936±0.038	0.961±0.008	0.948±0.020	0.983±0.003
100/20	—	T1	0.988±0.009	0.973±0.011	0.980±0.008	0.987±0.003
500/100	x	T1	0.982±0.007	0.964±0.011	0.972±0.005	0.979±0.005
500/100	—	T1	0.991±0.003	0.972±0.013	0.981±0.006	0.984±0.003
1000/200	x	T1	0.984±0.011	0.959±0.009	0.971±0.006	0.979±0.004
1000/200	—	T1	0.991±0.003	0.974±0.014	0.982±0.006	0.984±0.004

O treino com 1000 épocas e sem reinicialização dos pesos a cada *fold* obteve, de forma geral, as menores notas de avaliação deste bloco. Contudo, os outros parâmetros conseguiram medidas muito próximas, então os testes estatísticos foram realizados para verificar se existem diferenças significantes entre os parâmetros escolhidos.

Primeiro foi realizado o teste estatístico de *Friedman* para o *recall*, e como pode ser visto na Tabela 5.10 o valor de $\chi^2 = 37,29$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H_0) para 0,5 de significância. Adicionalmente pode ser observado que o *p-value* foi baixo, o que também indica que é possível rejeitar H_0 , ou seja, houve entre os experimentos diferenças estatisticamente

significativas para o *recall*.

Tabela 5.10: Valores estatísticos do teste de *Friedman* para o *recall* do bloco de experimentos 2.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	p-value
Teste	37,29	0,005
H0 para 0,5	11,07	

Para verificar quais experimentos contribuíram para essas diferenças foi feito uma análise *post hoc* pelo método comparativo de *Bonferroni*, seus resultados podem ser verificados na Tabela 5.11 e na Figura 5.4.

Tabela 5.11: Valores do método de *Bonferroni* para o *recall* do bloco de experimentos 2.

Experimentos	L. Inferior	Média	L. Superior	P-Value
1 2	-6,402	-4,05	-1,698	0,00001
1 3	-3,802	-1,45	0,902	1,000
1 4	-5,902	-3,55	-1,198	0,000
1 5	-4,752	-2,4	-0,048	0,041
1 6	-5,902	-3,55	-1,198	0,000
2 3	0,248	2,6	4,952	0,018
2 4	-1,852	0,5	2,852	1,000
2 5	-0,702	1,65	4,002	0,592
2 6	-1,852	0,5	2,852	1,000
3 4	-4,452	-2,1	0,252	0,132
3 5	-3,302	-0,95	1,402	1,000
3 6	-4,452	-2,1	0,252	0,132
4 5	-1,202	1,15	3,502	1,000
4 6	-2,352	0	2,352	1,000
5 6	-3,502	-1,15	1,202	1,000

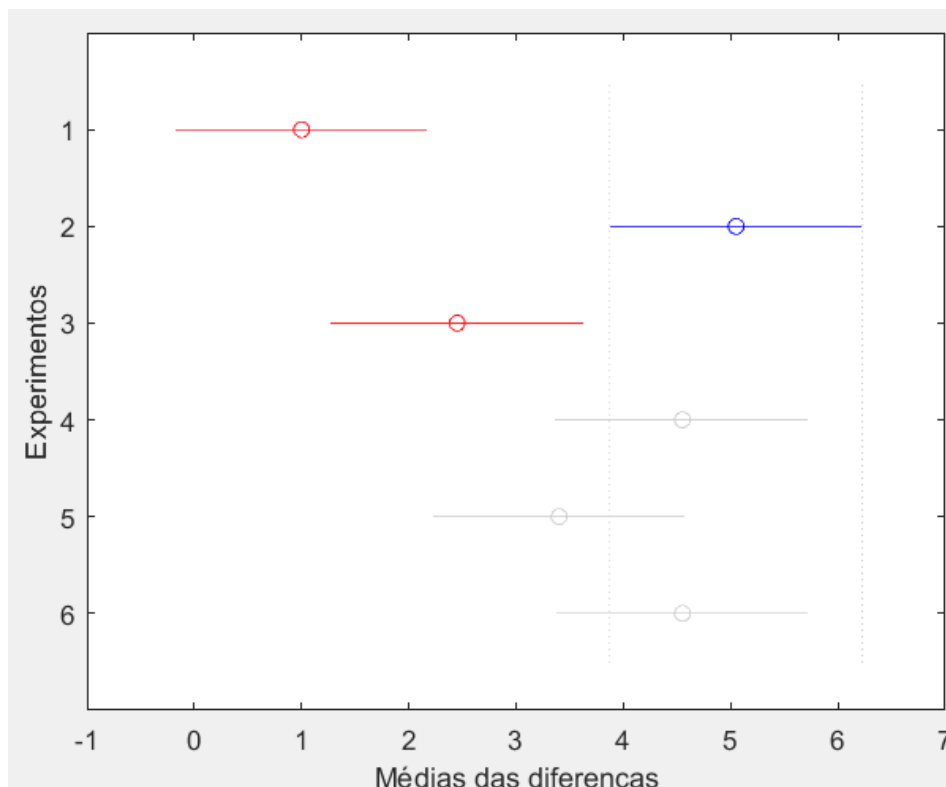


Figura 5.4: Gráfico do método de *Bonferroni* para o *recall* do bloco de experimentos 2.

Da análise por *Bonferroni*, conforme exibido na Tabela 5.11 e na Figura 5.4, é possível verificar que os experimentos 2,4,5 e 6 não possuem diferenças significativas em seus resultados, porém os parâmetros utilizados nos experimentos 1 e 2 foram os que menos contribuíram para a melhoria do *recall*.

A próxima medida a ser verificada é a precisão, e na Tabela 5.12 é visto as estatísticas do teste de *Friedman*. O valor de $\chi^2 = 27,74$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H_0) para 0,5 de significância. Logo, é possível rejeitar H_0 , ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas de precisão.

Tabela 5.12: Valores estatísticos do teste de *Friedman* para a precisão do bloco de experimentos 2.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	p-value
Teste	27,74	0,40
H_0 para 0,5	11,07	

Visto que existem diferenças significativas uma análise *post hoc* por *Bonferroni* foi executada para verificar quais parâmetros nos experimentos contribuíram para essa variação. A Tabela 5.13 exhibe as comparações múltiplas entre os experimentos para este teste e na Figura 5.5 é possível visualizar suas variações.

Tabela 5.13: Valores do teste de *Bonferroni* para a precisão do bloco de experimentos 2.

Experimentos	L. Inferior	Média	L. Superior	<i>P-Value</i>
1 2	-5,345	-2,9	-0,455	0,007
1 3	-2,895	-0,45	1,995	1,000
1 4	-4,645	-2,2	0,245	0,124
1 5	-2,195	0,25	2,695	1,000
1 6	-4,945	-2,5	-0,055	0,040
2 3	0,005	2,45	4,895	0,049
2 4	-1,745	0,7	3,145	1,000
2 5	0,705	3,15	5,595	0,002
2 6	-2,045	0,4	2,845	1,000
3 4	-4,195	-1,75	0,695	0,535
3 5	-1,745	0,7	3,145	1,000
3 6	-4,495	-2,05	0,395	0,208
4 5	0,005	2,45	4,895	0,049
4 6	-2,745	-0,3	2,145	1,000
5 6	-5,195	-2,75	-0,305	0,014

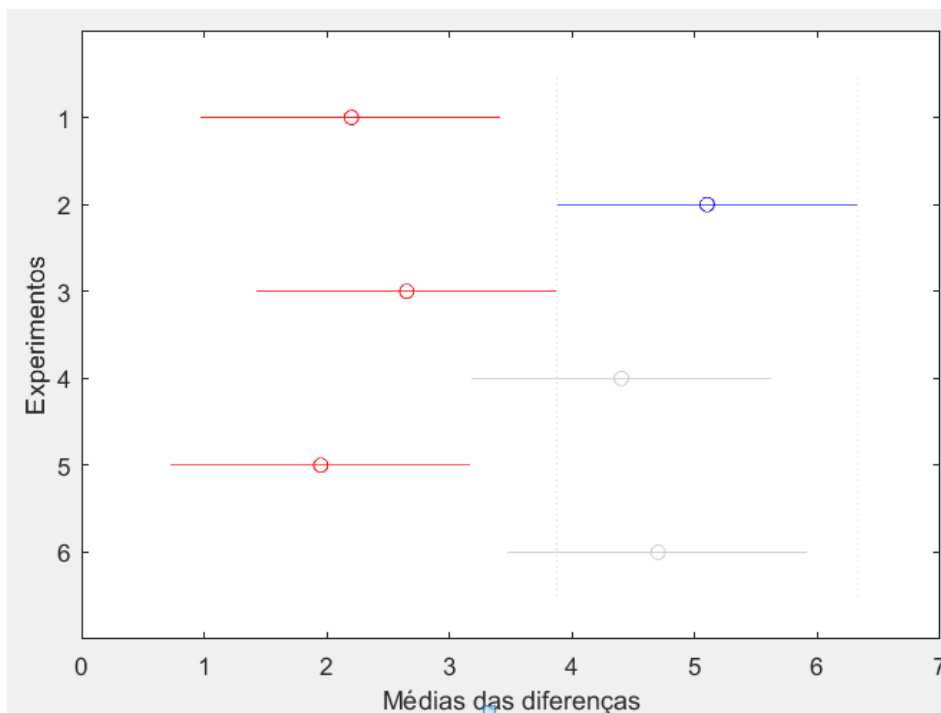


Figura 5.5: Gráfico do método de *Bonferroni* para a precisão do bloco de experimentos 2.

Os experimentos 2,4 e 6 não possuem diferenças significativas para a precisão e foram os experimentos que neste quesito obtiveram as melhores medidas. Os experimentos 1, 3 e 5 foram os que contribuíram menos para a melhoria da precisão.

Agora serão exibidos na Tabela 5.14 os resultados do teste de *Friedman* para a acurácia.

Tabela 5.14: Valores estatísticos do teste de *Friedman* para a acurácia do bloco de experimentos 2.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	p-value
Teste	39,39	0,002
H0 para 0,5	11,07	

Conforme exibido na Tabela 5.14, o valor de $\chi^2 = 39,39$ e este é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno (0,002) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas de acurácia.

Para verificar quais experimentos contribuíram para essas diferenças, foi feita uma análise comparativa do tipo *Bonferroni*, seus resultados podem ser verificados na Tabela 5.15 e na Figura 5.6.

Tabela 5.15: Valores do método de *Bonferroni* para a acurácia do bloco de experimentos 2.

Experimentos	L. Inferior	Média	L. Superior	<i>P-Value</i>
1 2	-6,253	-3,85	-1,447	0,00004
1 3	-4,103	-1,7	0,703	0,567
1 4	-6,103	-3,7	-1,297	0,000
1 5	-4,103	-1,7	0,703	0,567
1 6	-6,453	-4,05	-1,647	0,000
2 3	-0,253	2,15	4,553	0,129
2 4	-2,253	0,15	2,553	1,000
2 5	-0,253	2,15	4,553	0,129
2 6	-2,603	-0,2	2,203	1,000
3 4	-4,403	-2	0,403	0,218
3 5	-2,403	0	2,403	1,000
3 6	-4,753	-2,35	0,053	0,061
4 5	-0,403	2	4,403	0,218
4 6	-2,753	-0,35	2,053	1,000
5 6	-4,753	-2,35	0,053	0,061

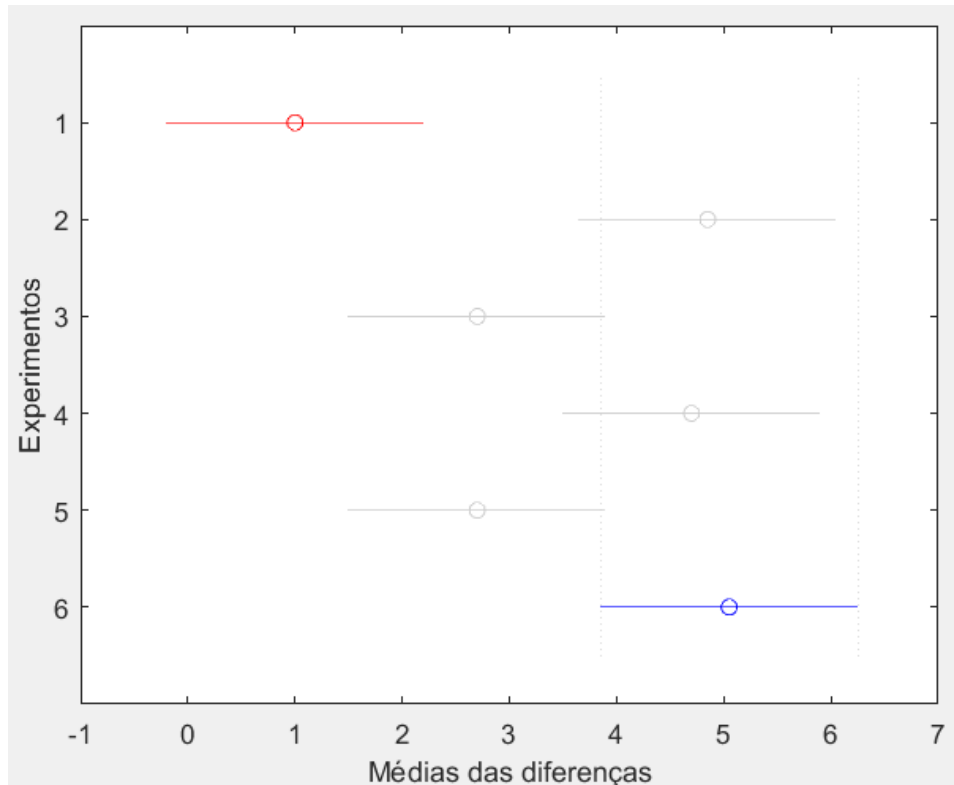


Figura 5.6: Gráfico do método de *Bonferroni* para a acurácia do bloco de experimentos 2.

O experimento 1 foi o que menos contribuiu para melhoria da acurácia, sendo estatisticamente semelhante aos experimentos 3 e 5. E o experimento 6 foi o que mais contribuiu para a melhoria da acurácia e ele é estatisticamente diferente do experimento 1 e semelhante aos experimentos 2,3,4 e 5.

Para finalizar os testes estatísticos deste bloco foi realizado o teste de *Friedman* para verificar se existem diferenças significativas entre os experimentos para a AUC conforme a Tabela 5.16.

Tabela 5.16: Valores estatísticos do teste de *Friedman* para a AUC do bloco de experimentos 2.

Nº de Algoritmos=6 Nº de domínios=10	χ^2	<i>p-value</i>
Teste	24,06	0,0002
H0 para 0,5	11,07	

Conforme exibido na Tabela 5.16, o valor de $\chi^2 = 24,06$ e este é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno (0,0002) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para a AUC.

Para verificar quais experimentos contribuíram para essas diferenças foi feito uma análise comparativa do tipo *Bonferroni*, seus resultados podem ser verificados na Tabela 5.17 e na Figura 5.7.

Tabela 5.17: Valores do método de *Bonferroni* para a AUC do bloco de experimentos 2.

Experimentos	L. Inferior	Média	L. Superior	<i>p-value</i>
1 2	-4,156	-1,7	0,756	0,632
1 3	-0,756	1,7	4,156	0,632
1 4	-2,756	-0,3	2,156	1,000
1 5	-0,956	1,5	3,956	1,000
1 6	-3,056	-0,6	1,856	1,000
2 3	0,944	3,4	5,856	0,001
2 4	-1,056	1,4	3,856	1,000
2 5	0,744	3,2	5,656	0,002
2 6	-1,356	1,1	3,556	1,000
3 4	-4,456	-2	0,456	0,252
3 5	-2,656	-0,2	2,256	1,000
3 6	-4,756	-2,3	0,156	0,090
4 5	-0,656	1,8	4,256	0,472
4 6	-2,756	-0,3	2,156	1,000
5 6	-4,556	-2,1	0,356	0,181

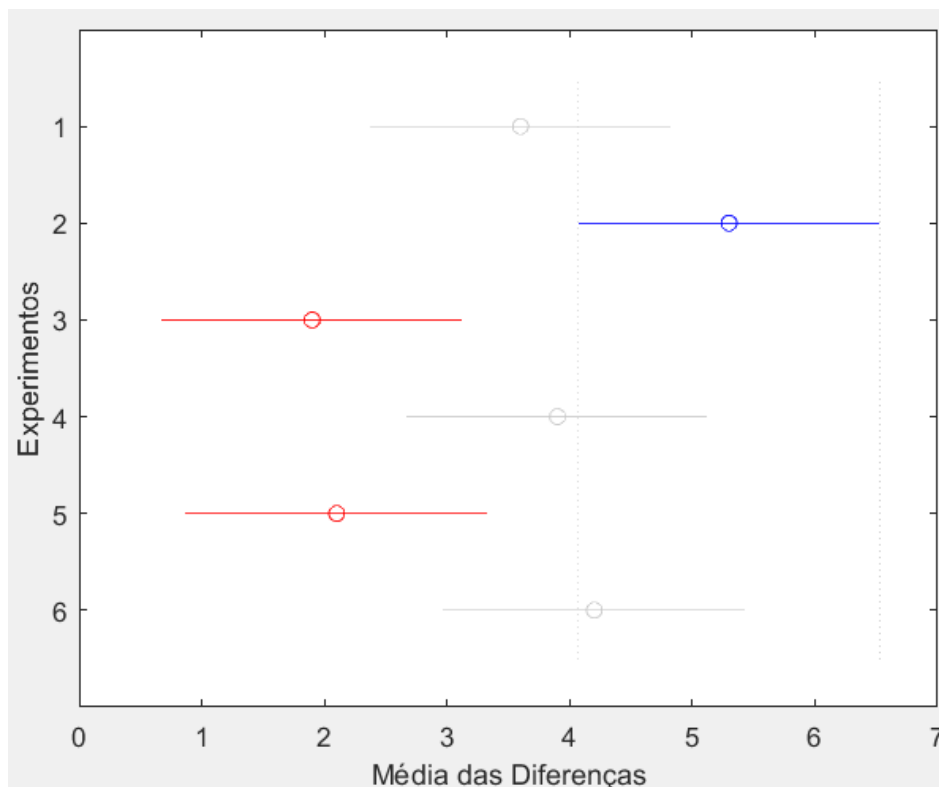


Figura 5.7: Gráfico do método de *Bonferroni* para a AUC do bloco de experimentos 2.

De acordo as análises, os experimentos que mais contribuíram para a melhoria da AUC foram 1, 2, 4 e 6 e os que menos contribuíram foram os experimentos 3 e 5.

De forma geral após os testes estatísticos e as devidas análises foi possível ver que o treinamento sem a reinicialização dos pesos contribuiu mais para a melhoria das medidas, porém na maioria dos casos não houve diferenças significativas para a quantidade de épocas utilizadas no treinamento, como o treino com 100 épocas pode levar a uma parada do treino antes de atingir uma boa generalização e existem mecanismos para evitar um sobre-ajuste sobre os dados de treino, o treino com 1000 épocas foi escolhido, logo estes foram os parâmetros para o próximo bloco de experimentos.

Nos dois primeiros blocos foram avaliados os parâmetros gerais de treinamento como: algoritmo de treino, quantidade de neurônios na camada intermediária, quantidade de épocas de treinamento e reinicialização de pesos. Agora será avaliado o comportamento da rede a cada variação dos preditores.

Os preditores foram divididos em três grupos (G1, G2 e G3) para que assim fosse avaliado o impacto de cada grupo individualmente e a interação destes na capacidade de classificação da rede conforme os valores exibidos na Tabela 5.18.

Tabela 5.18: Medidas de avaliação dos parâmetros para o bloco de experimentos 3.

G	G	G	Test	<i>Recall</i>	Precisão	Acurácia	AUC
1	2	3	e	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$
x	-	-	T1	0.958±0.052	0.963±0.017	0.961±0.031	0.979±0.005
-	x	-	T1	0.514±0.025	0.844±0.019	0.709±0.013	0.833±0.013
-	-	x	T1	0.900±0.012	0.944±0.013	0.923±0.005	0.971±0.004
-	x	x	T1	0.910±0.010	0.961±0.011	0.936±0.003	0.986±0.002
x	x	-	T1	0.988±0.003	0.976±0.009	0.982±0.004	0.981±0.004
x	-	x	T1	0.987±0.004	0.970±0.009	0.978±0.004	0.980±0.004
x	x	x	T1	0.991±0.003	0.974±0.014	0.982±0.006	0.984±0.004

Analisando a Tabela 5.18 observa-se que a combinação dos grupos G1 e G2 alcançou o segundo melhor resultado deste bloco e o grupo G2 sozinho alcançou os resultados mais baixos. Mas é possível notar que a avaliação da combinação dos três grupos ainda obteve uma avaliação mais alta no geral, porém é necessário realizar os testes estatísticos para verificar se as diferenças entre os experimentos têm significância estatística.

Na Tabela 5.19 é exibido as estatísticas do teste de *Friedman* para o *recall* do Bloco 3. O valor do $\chi^2 = 53,03$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E observando o *p-value* que foi bem pequeno (0,00001) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para o *recall*.

Tabela 5.19: Valores estatísticos do teste de *Friedman* para o *recall* do bloco de experimentos 3.

Nº de Algoritmos=7 Nº de domínios=10	χ^2	p-value
Teste	53,03	0,00001
H0 para 0,5	11,07	

Visto que existem diferenças significativas uma análise por *Bonferroni* foi executada para verificar quais parâmetros nos experimentos contribuíram para essa variação. A Tabela 5.20 exibe as comparações múltiplas entre os experimentos para esta análise e na Figura 5.8 é possível visualizar suas variações.

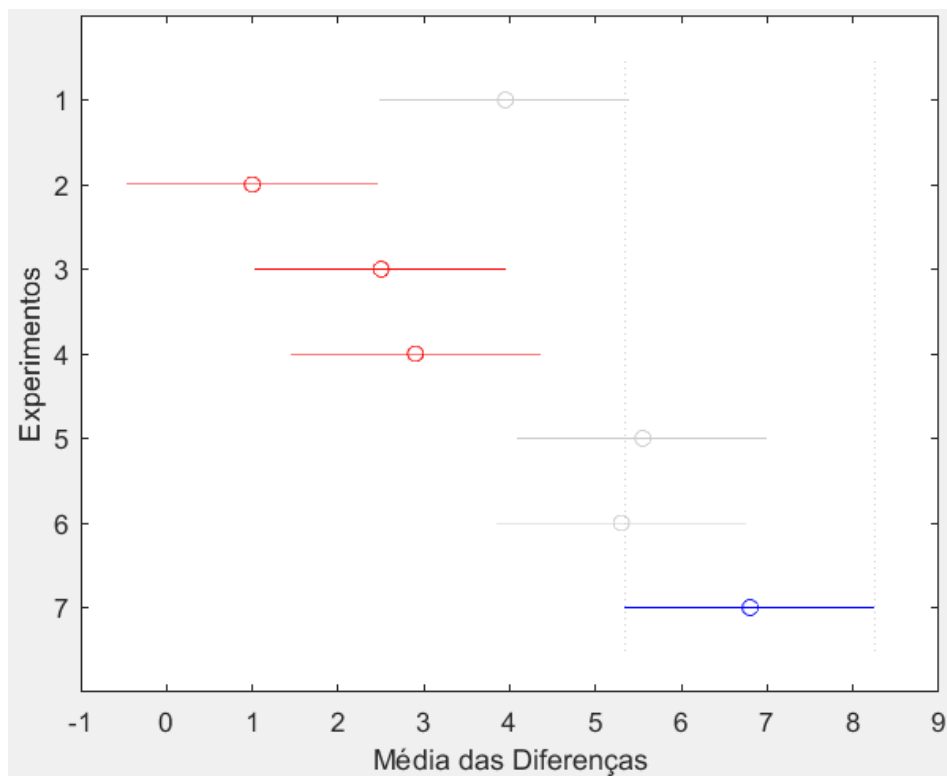


Figura 5.8: Gráfico do método de *Bonferroni* para o *recall* do bloco de experimentos 3.

Tabela 5.20: Valores estatísticos do método de *Bonferroni* para o *recall* do bloco de experimentos 3.

Experimentos		L. Inferior	Média	L. Superior	<i>P-Value</i>
1	2	0,036	2,95	5,864	0,044
1	3	-1,464	1,45	4,364	1,000
1	4	-1,864	1,05	3,964	1,000
1	5	-4,514	-1,6	1,314	1,000
1	6	-4,264	-1,35	1,564	1,000
1	7	-5,764	-2,85	0,064	0,062
2	3	-4,414	-1,5	1,414	1,000
2	4	-4,814	-1,9	1,014	1,000
2	5	-7,464	-4,55	-1,636	0,000
2	6	-7,214	-4,3	-1,386	0,000
2	7	-8,714	-5,8	-2,886	0,000
3	4	-3,314	-0,4	2,514	1,000
3	5	-5,964	-3,05	-0,136	0,031
3	6	-5,714	-2,8	0,114	0,074
3	7	-7,214	-4,3	-1,386	0,000
4	5	-5,564	-2,650	0,264	0,120
4	6	-5,314	-2,400	0,514	0,259
4	7	-6,814	-3,900	-0,986	0,001
5	6	-2,664	0,250	3,164	1,000
5	7	-4,164	-1,250	1,664	1,000
6	7	-4,414	-1,500	1,414	1,000

Observando os resultados da análise para o *recall* no Bloco 3 e pela Figura 5.8 é possível perceber que os parâmetros do experimento 7 promoveram variações positivas no *recall* estatisticamente semelhantes aos experimentos 1, 5 e 6, porém

diferentes dos experimentos 2, 3 e 4 que contribuíram menos para a melhoria desta medida.

Para verificar os efeitos dos parâmetros do Bloco 3 na precisão os testes estatísticos foram realizados e exibidos na Tabela 5.21.

Tabela 5.21: Valores estatísticos do teste de *Friedman* para a precisão do Bloco de experimentos 3.

Nº de Algoritmos=7 Nº de domínios=10	χ^2	p-value
Teste	50,46	0,00003
H0 para 0,5	11,07	

Conforme exibido na Tabela 5.21, o valor de $\chi^2 = 50,46$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno (0,00003) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas de precisão.

Para verificar quais experimentos contribuíram para essas diferenças foi feita uma análise *post hoc* pelo método de *Bonferroni*, seus resultados podem ser verificados na Tabela 5.22 e na Figura 5.9.

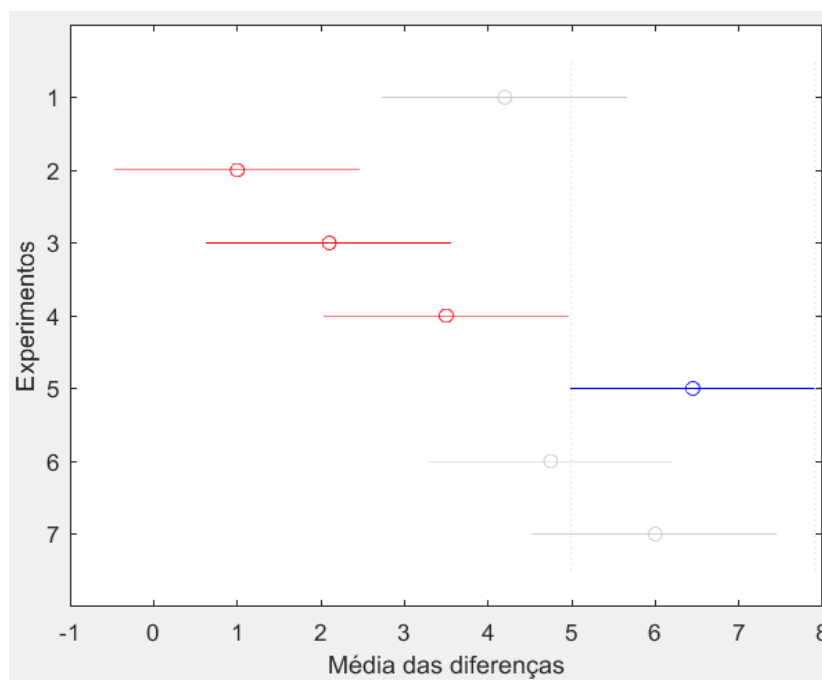


Figura 5.9: Gráfico do método de *Bonferroni* para a precisão do bloco de experimentos 3.

Tabela 5.22: Valores do método de *Bonferroni* para a precisão do Bloco de experimentos 3.

Experimentos		L. Inferior	Média	L. Superior	<i>P-Value</i>
1	2	0,270	3,2	6,130	0,019
1	3	-0,830	2,1	5,030	0,618
1	4	-2,230	0,7	3,630	1,000
1	5	-5,180	-2,25	0,680	0,412
1	6	-3,480	-0,55	2,380	1,000
1	7	-4,730	-1,8	1,130	1,000
2	3	-4,030	-1,1	1,830	1,000
2	4	-5,430	-2,5	0,430	0,200
2	5	-8,380	-5,45	-2,520	0,000
2	6	-6,680	-3,75	-0,820	0,002
2	7	-7,930	-5	-2,070	0,000
3	4	-4,330	-1,4	1,530	1,000
3	5	-7,280	-4,35	-1,420	0,000
3	6	-5,580	-2,65	0,280	0,126
3	7	-6,830	-3,9	-0,970	0,001
4	5	-5,880	-2,950	-0,020	0,047
4	6	-4,180	-1,250	1,680	1,000
4	7	-5,430	-2,500	0,430	0,200
5	6	-1,230	1,700	4,630	1,000
5	7	-2,480	0,450	3,380	1,000
6	7	-4,180	-1,250	1,680	1,000

Apesar do destaque do experimento 5 sobre a precisão, ele é estatisticamente semelhante aos experimentos 1, 6 e 7, porém contribui mais para o aumento desta medida do que os experimentos 2, 3 e 4.

O próximo teste estatístico foi aplicado para verificar se existem diferenças entre os experimentos para a acurácia, e na Tabela 5.23 é exibido o resultado.

Tabela 5.23: Valores estatísticos do teste de *Friedman* para a acurácia do bloco de experimentos 3.

Nº de Algoritmos=7 Nº de domínios=10	χ^2	p-value
Teste	52,72	0,00001
H0 para 0,5	11,07	

Conforme exibido na Tabela 5.23, o valor de $\chi^2 = 52,72$ é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno (0,00001) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para as medidas de acurácia.

Para verificar quais experimentos contribuíram para essas diferenças foi feita uma análise comparativa usando *Bonferroni*, seus resultados podem ser verificados na Tabela 5.24 e na Figura 5.10.

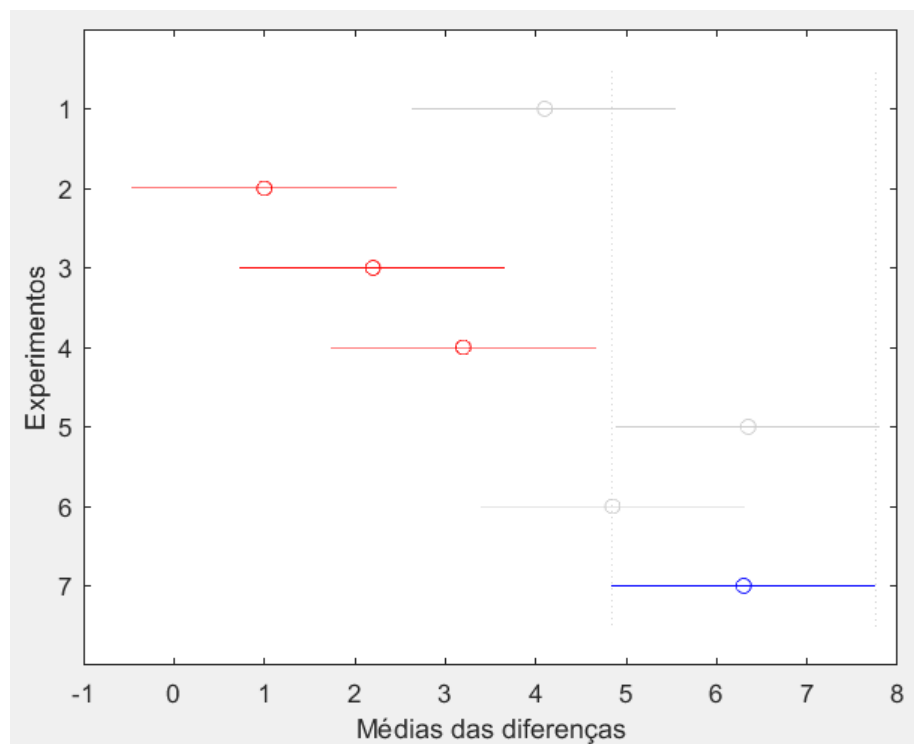


Figura 5.10: Gráfico do método de *Bonferroni* para a acurácia do bloco de experimentos 3.

Tabela 5.24: Valores do método de *Bonferroni* para a acurácia do bloco de experimentos 3.

Experimentos	L. Inferior	Média	L. Superior	<i>P-value</i>
1 2	0,175	3,1	6,025	0,027
1 3	-1,025	1,9	4,825	1,000
1 4	-2,025	0,9	3,825	1,000
1 5	-5,175	-2,25	0,675	0,408
1 6	-3,675	-0,75	2,175	1,000
1 7	-5,125	-2,2	0,725	0,468
2 3	-4,125	-1,2	1,725	1,000
2 4	-5,125	-2,2	0,725	0,468
2 5	-8,275	-5,35	-2,425	0,000
2 6	-6,775	-3,85	-0,925	0,001
2 7	-8,225	-5,3	-2,375	0,000
3 4	-3,925	-1	1,925	1,000
3 5	-7,075	-4,15	-1,225	0,000
3 6	-5,575	-2,65	0,275	0,124
3 7	-7,025	-4,1	-1,175	0,000
4 5	-6,075	-3,150	-0,225	0,022
4 6	-4,575	-1,650	1,275	1,000
4 7	-6,025	-3,100	-0,175	0,027
5 6	-1,425	1,500	4,425	1,000
5 7	-2,875	0,050	2,975	1,000
6 7	-4,375	-1,450	1,475	1,000

Pela análise realizada cujos resultados foram exibidos na Tabela 5.24 e na Figura 5.10 é possível verificar que os experimentos 1, 5, 6 e 7 não possuem diferenças significativas para a acurácia, porém os experimentos 5 e 7 contribuíram mais para a melhoria desta medida do que os experimentos 2, 3 e 4.

Por fim, foram verificadas se existem diferenças significativas entre os experimentos do Bloco 3 para a AUC. O resultado do teste de *Friedman* é exibido na Tabela 5.25.

Tabela 5.25: Valores estatísticos do teste de *Friedman* para a AUC do bloco de experimentos 3.

Nº de Algoritmos=7 Nº de domínios=10	χ^2	p-value
Teste	47,87	0,0001
H0 para 0,5	11,07	

Conforme exibido na Tabela 5.25, o valor de $\chi^2 = 47,87$ e este é maior do que o valor crítico (11,07) para rejeitar a hipótese nula (H0) para 0,5 de significância. E ainda observando o *p-value* que foi bem pequeno (0,0001) é possível rejeitar H0, ou seja, existem diferenças estatisticamente significativas nos experimentos para a AUC.

Para verificar quais experimentos contribuíram para essas diferenças foi realizada a análise comparativa usando *Bonferroni*, seus resultados podem ser verificados na Tabela 5.26 e na Figura 5.11.

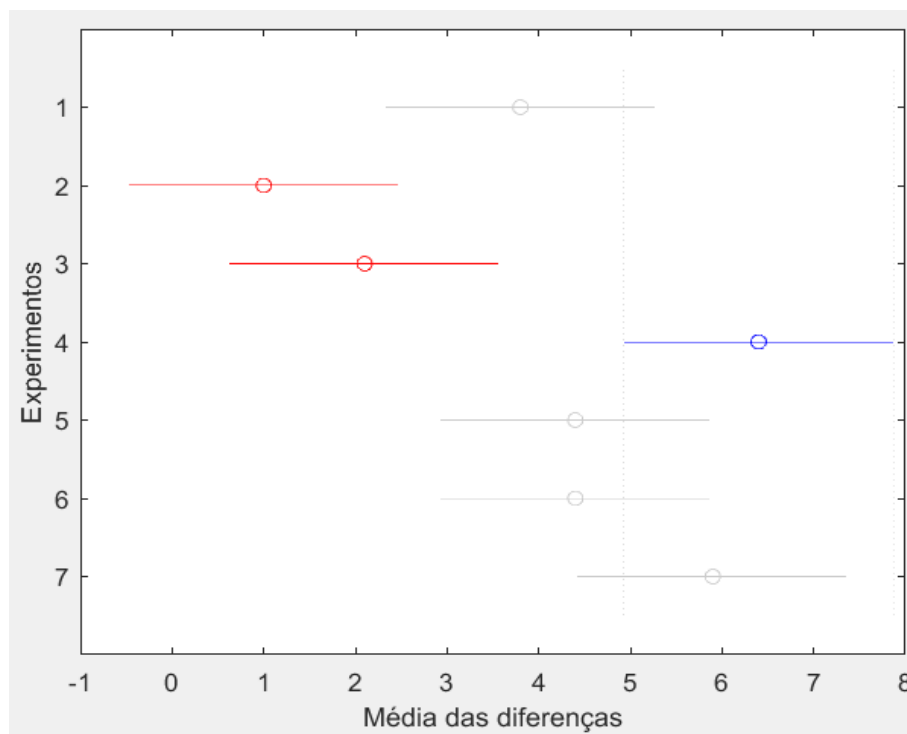


Figura 5.11: Gráfico do método de *Bonferroni* para a AUC do bloco de experimentos 3.

Tabela 5.26: Valores do método de *Bonferroni* para a AUC do bloco de experimentos 3.

Experimentos		L. Inferior	Média	L. Superior	<i>P-Value</i>
1	2	-0,135	2,8	5,735	0,079
1	3	-1,235	1,7	4,635	1,000
1	4	-5,535	-2,6	0,335	0,149
1	5	-3,535	-0,6	2,335	1,000
1	6	-3,535	-0,6	2,335	1,000
1	7	-5,035	-2,1	0,835	0,624
2	3	-4,035	-1,1	1,835	1,000
2	4	-8,335	-5,4	-2,465	0,000
2	5	-6,335	-3,4	-0,465	0,009
2	6	-6,335	-3,4	-0,465	0,009
2	7	-7,835	-4,9	-1,965	0,000
3	4	-7,235	-4,3	-1,365	0,000
3	5	-5,235	-2,3	0,635	0,363
3	6	-5,235	-2,3	0,635	0,363
3	7	-6,735	-3,8	-0,865	0,002
4	5	-0,935	2,000	4,935	0,807
4	6	-0,935	2,000	4,935	0,807
4	7	-2,435	0,500	3,435	1,000
5	6	-2,935	0,000	2,935	1,000
5	7	-4,435	-1,500	1,435	1,000
6	7	-4,435	-1,500	1,435	1,000

De acordo as análises os experimentos que mais contribuíram para a melhoria da AUC foram 1, 4, 5,6 e 7 e os que menos contribuíram foram os experimentos 2 e 3.

De forma geral após os testes estatísticos foi possível ver que o treinamento com todos os grupos de preditores contribuiu mais para a melhoria das medidas,

porém na maioria dos casos os experimentos 2 (só com o grupo 2), 3 (só com o grupo 3) e 4 (grupos 2 e 3 juntos) foram os que possuíram as medidas mais baixas.

Após os testes estatísticos para os três blocos de experimentos, os parâmetros de treinamento que se sobressaíram produziram uma RNA treinada com o algoritmo LMB, com 1000 épocas de treino e 200 de validação, 26 neurônios na camada de entrada correspondente aos 26 preditores, com 10 neurônios na camada intermediária e sem reinicialização dos pesos a cada *fold*.

Uma vez então escolhidos os parâmetros de treinamento, alguns testes foram realizados com as amostras de testes T2 e T3 para avaliar a classificação da rede para os espectros de duas galáxias previamente escolhidas e ainda avaliar o desempenho do método em espectros de galáxias com resolução espectral mais baixa do que os espectros utilizados no treinamento.

5.2 Treino e Testes Finais

A amostra de testes T1 foi selecionada aleatoriamente a cada *fold* representando assim 10% dos dados de treino, pois foram utilizadas 10 *folds*. Esta amostra conteve porções diversas dos variados espectros utilizados no treino. Apesar de dar uma idéia geral da capacidade de generalização da Rede Neural para esta amostra, ainda é necessário saber como serão os resultados para um espectro completo, tal qual serão utilizados pelos futuros usuários desta abordagem aqui desenvolvida.

Então, para uma verificação adicional da abordagem de treino com melhor avaliação até o momento, foi realizado um teste com duas amostras distintas, a primeira amostra T1 contendo 10% dos dados de treino a cada *fold* e a segunda amostra T2 que contém espectros vindos do mesmo instrumento utilizado para adquirir os dados de treino. As medidas de avaliação para as amostras de teste T1, e T2 podem ser visualizadas na Tabela 5.27.

Tabela 5.27: Medidas de avaliação para as amostras de teste T1 e T2.

Teste	<i>Recall</i>	Precisão	Acurácia	AUC
	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$	$\underline{x} \pm S$
T1	0.991±0.003	0.974±0.014	0.982±0.006	0.984±0.005
T2	0.927±0.008	0.721±0.011	0.972±0.001	0.964±0.0009

As medidas verificadas na Tabela 5.27 são obtidas pela média das medidas tomadas da classificação da rede treinada em cada *fold* do *cross-validation*. A utilização do *cross-validation* até agora serviu para dar uma estimativa geral da capacidade de generalização da RNA para dados desconhecidos do treinamento. Esta estimativa pôde ser comprovada pelos resultados obtidos com a amostra de testes T2. Porém, como deseja-se um classificador que possa ser utilizado em uma rotina computacional, uma só Rede Neural foi treinada com todos os dados de treino e o resultado pode ser visto na Tabela 5.28 onde são exibidas as medidas para a classificação desta rede para a amostra de teste T2 e para a amostra T3, sendo esta última obtida de outro instrumento com resolução espectral mais baixa e uma taxa de sinal ruído diferente.

Tabela 5.28: Medidas de avaliação para as amostras de teste T2 e T3.

<i>Teste</i>	<i>Recall</i>	Precisão	Acurácia	AUC
T2	0.932	0.687	0.968	0.981
T3	0.938	0.629	0.968	0.965

Neste novo treinamento a RNA alcançou um *recall* melhor para a amostra T2 do que durante o treino com *cross-validation*. Este *recall* foi semelhante para a amostra T3. Porém, é importante ressaltar que a amostra T3 é constituída por espectros obtidos em um instrumento diferente do utilizado nos espectros do treino e que possui uma resolução um pouco mais baixa (5,9 Å) do que a amostra de treino (5,6 Å), contudo foi alcançado um *recall* de 0,938 e uma acurácia de 0,968. Um dos

fatores que levam a precisão nos testes T2 e T3 ser mais baixa em relação ao teste T1, como exibido na Tabela 5.28, é o desbalanceamento entre a quantidade de linhas em emissão e ruído existentes nos espectros. No treino esse desbalanceamento foi compensado por meio de um processo de superamostragem das linhas em emissão para que se igualasse em quantidade ao ruído, porém nas amostras do teste final essa superamostragem não pode ser realizada, já que não se conhece quais são as linhas existentes no espectro em uma situação real.

Para a amostra de teste T3 as linhas em emissão representam apenas 4,9% das classes e para a amostra T2 estas linhas representam 6,3%. Contudo, o classificador reconheceu corretamente mais de 93% de ambas as classes. Uma outra maneira de visualizar o desempenho do classificador para ambas as classes é por meio da matriz de confusão como mostram as Tabelas 5.29 e 5.30.

Tabela 5.29: Matriz de confusão para amostra de teste T2.

Matriz de Confusão		
Classes	Linha	Ruído
Esperado Linha	55 93,2%	25 9,1%
Esperado Ruído	4 6,8%	841 90,9%
Total	59 100%	866 100%

Real

Tabela 5.30: Matriz de confusão para amostra de teste T3.

Matriz de Confusão		
Classes	Linha	Ruído
Esperado Linha	44 93,6%	26 3,0%
Esperado Ruído	3 6,4%	829 97,0%
Total	81 100%	855 100%

Real

Com a matriz de confusão é possível ver na diagonal principal que, tanto na amostra de T2 quanto em T3, a RNA classificou corretamente mais de 90% de ambas as classes, e na diagonal secundária é visto que o erro ficou abaixo de 10%, mas além da matriz de confusão, outra forma de visualizar as classificações obtidas pela rede é por meio da Curva ROC exibida nas Figura 5.12 e Figura 5.13.

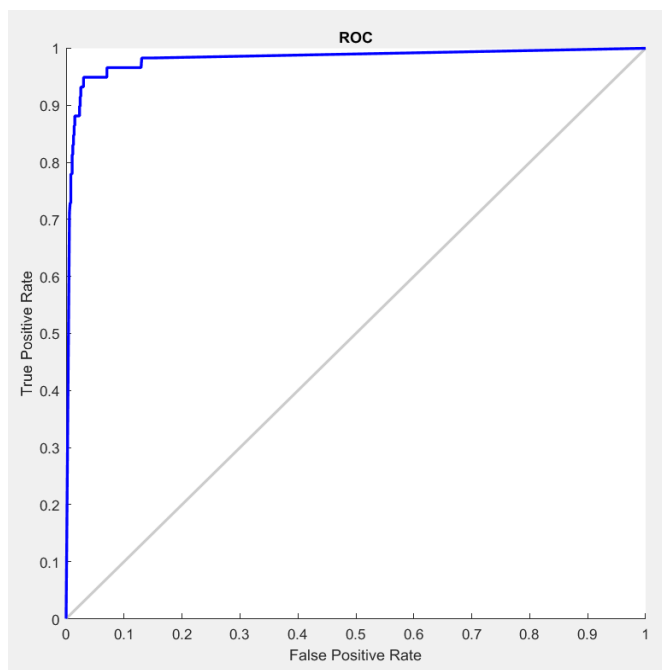


Figura 5.12: Curva ROC para a amostra de teste T2.

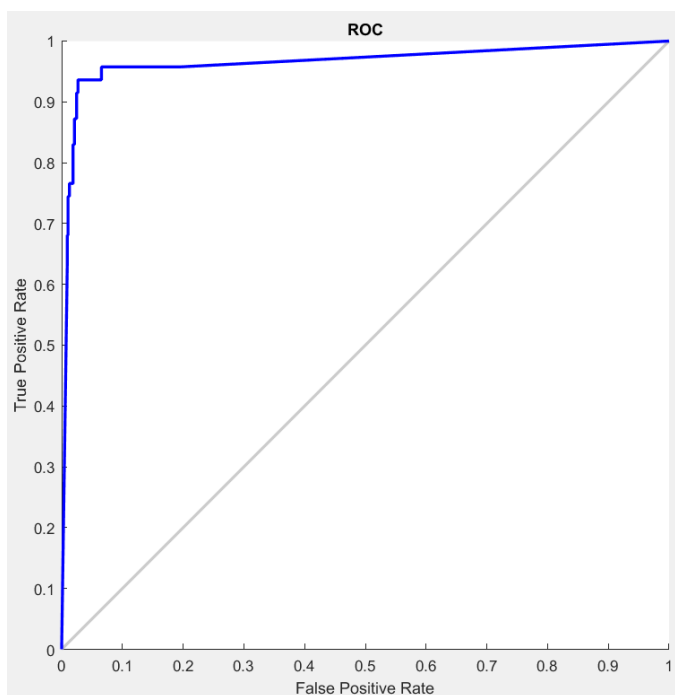


Figura 5.13: Curva ROC para a amostra de teste T3.

É possível ver pelas curvas ROC que a RNA consegue resultados consistentes para ambas as amostras, porém, seu desempenho é um pouco melhor para espectros com resolução espectral semelhante aos utilizados no treino.

Após a visualização dos resultados por meio dos gráficos e tabelas, é ainda importante verificar a classificação da RNA diretamente nos espectros testados e as Figura 5.14 e 5.15 exibem estas classificações para a amostra T3.

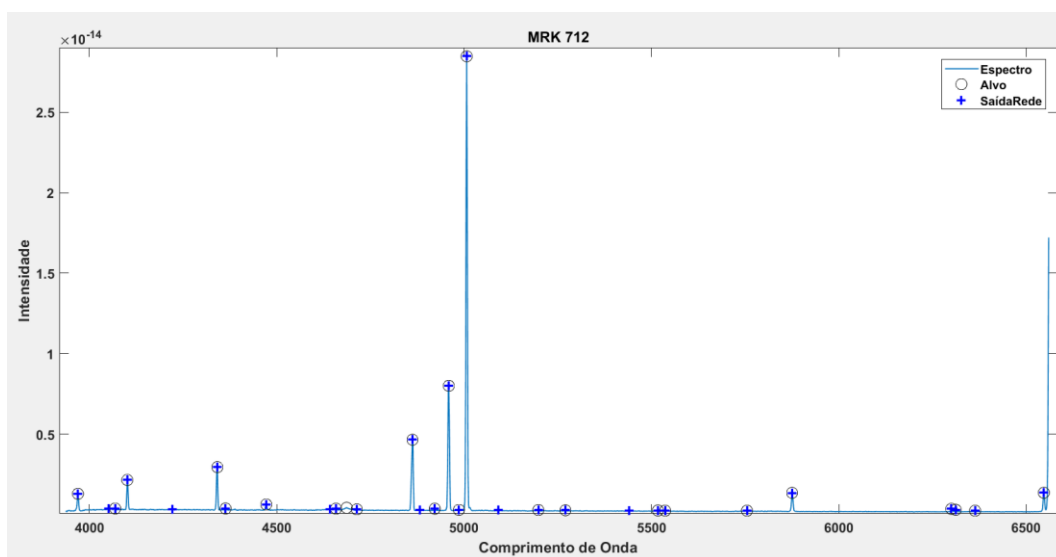


Figura 5.14: Espectro da galáxia MRK 712 com classificações da RNA.

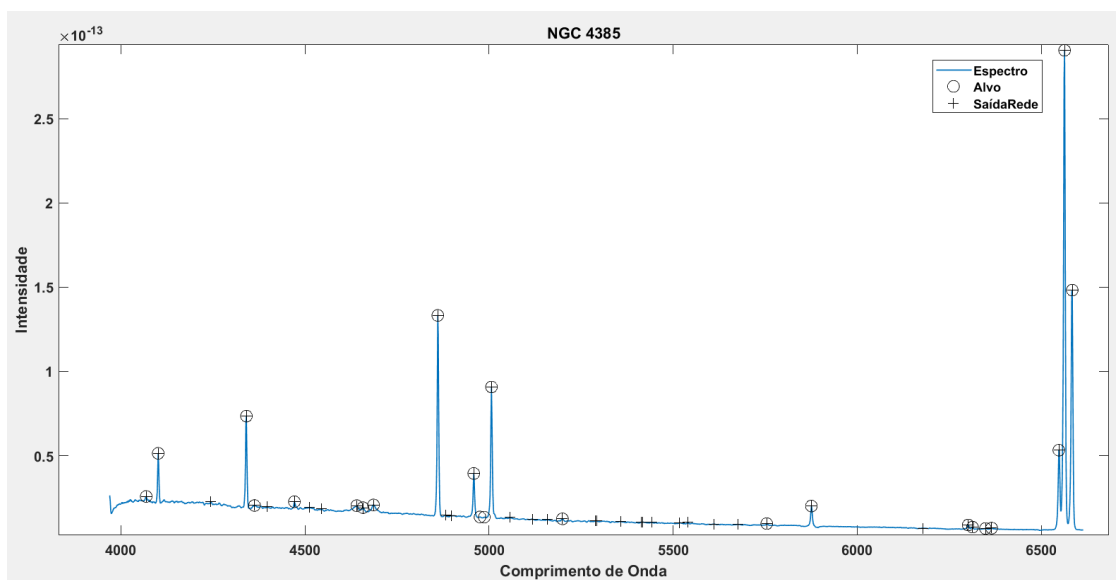


Figura 5.15: Espectro da galáxia NGC 4385 com classificações da RNA.

Analisando as classificações é possível verificar que a abordagem desenvolvida nesta pesquisa consegue encontrar determinadas linhas mesmo que estas estejam

imersas no ruído e com seu perfil alterado, como a linha 4685.71 HeII exibida ao centro na Figura 5.16, que é o resultado da sobreposição de três espectros de galáxias diferentes usadas no teste.

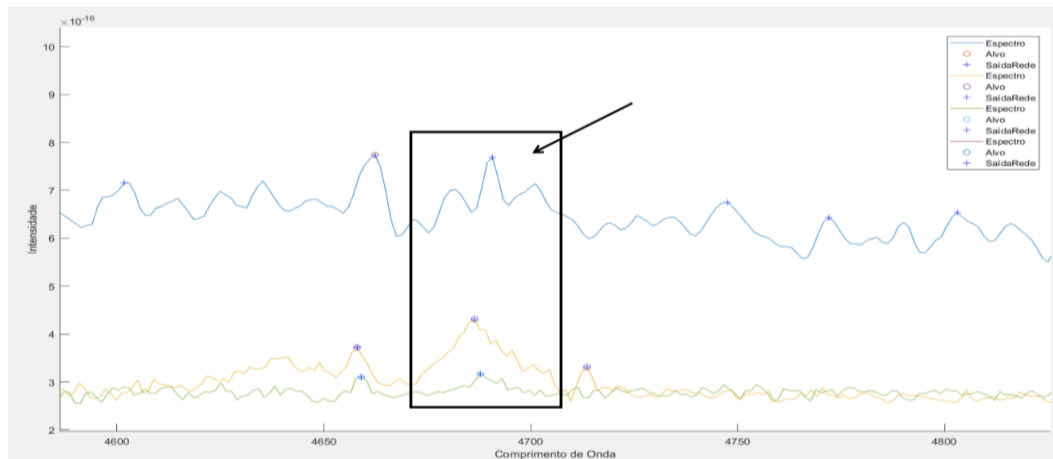


Figura 5.16: Sobreposição de linhas espectrais de três galáxias.

Como visto na Figura 5.16, a linha, localizada no centro e acima, não foi identificada pelo especialista na classificação prévia, provavelmente devido a sua similaridade com o ruído, porém foi encontrada pela abordagem. Esta figura também demonstra a variedade de formas que uma linha em emissão pode assumir em cada espectro, o que dificulta tanto a correta identificação por especialistas, como também por rotinas automáticas.

No entanto, a rotina computacional aqui desenvolvida tem conseguido contornar este problema, bem como o problema da identificação de linhas fracas, pois foram classificadas corretamente 100% das linhas fortes e 89% das linhas fracas das amostras de testes.

A RNA obtida neste treinamento final foi incorporada a uma rotina computacional que ao receber um espectro unidimensional como entrada, devolve como saída uma tabela com as linhas em emissão identificadas e suas respectivas medidas de largura e elevação. Além disso, exibe um gráfico do espectro com as devidas linhas.

5.3 Consistência Externa

Nesta seção os resultados da classificação obtida pela RNA, testada anteriormente, estão sendo comparados com as classificações de outros algoritmos de aprendizado de máquina. Esta classificação é realizada nas mesmas amostras de testes T2 e T3 utilizadas anteriormente e os algoritmos estão sendo treinados com a mesma amostra de treino usada pela RNA. Após isto foi feita uma comparação entre as linhas identificadas pela rotina computacional desenvolvida por meio desta pesquisa e pelo ALFA (WESSON, 2016).

Os algoritmos testados foram: Regressão Logística (RL) e Máquina de Vetor de Suporte (SVM), as medidas para as classificações podem ser vistas na Tabela 5.31.

Tabela 5.31: Comparação de medidas dos classificadores testados para duas amostras de testes.

Algoritmo	Teste	<i>Recall</i>	Precisão	Acurácia	AUC
RL	T2	0.898	0.449	0.923	0.971
	T3	0.893	0.344	0.905	0.949
SVM	T2	0.915	0.473	0.929	0.922
	T3	0.914	0.405	0.925	0.920
RNA	T2	0.932	0.687	0.968	0.981
	T3	0.936	0.629	0.968	0.965

Na Tabela 5.31 pode-se comparar as medidas alcançadas pela RNA e as medidas dos outros classificadores. É possível verificar que a RNA obteve os valores mais altos em todas as medidas para os dois conjuntos de testes. Porém, é necessário verificar a significância estatística destes resultados.

Testes Estatísticos Para os Classificadores

Como a amostra T2 possui 4 espectros e a amostra T3 possui 2 espectros, os testes foram feitos para as 6 medidas realizadas. A Tabela 5.32 exhibe o resultado do teste de Friedman para o *recall*.

Tabela 5.32: Valores estatísticos do teste de *Friedman* para o *recall* dos classificadores.

Nº de Algoritmos=3 Nº de domínios=6	χ^2	p-value
Teste	4	0,135
H0 para 0,5	7	

Como o valor do χ^2 foi de apenas 4, não é suficiente para determinar uma diferença estatística entre os resultados, pois o valor crítico seria de 7 de acordo a Japkowicz (2011). Portanto, em termos de *recall* pode-se dizer que os três classificadores demonstram o mesmo desempenho para os testes executados.

Em seguida, na Tabela 5.33 é exibido o teste de *Friedman* para a precisão.

Tabela 5.33: Valores estatísticos do teste de *Friedman* para a precisão dos classificadores.

Nº de Algoritmos=3 Nº de domínios=6	χ^2	p-value
Teste	9,33	0,009
H0 para 0,5	7	

O valor do χ^2 para a precisão foi de 9,33, superior ao valor crítico 7, o que significa que existe diferença estatística para a precisão. Então uma análise *post hoc* utilizando *Bonferroni* foi executada para verificar qual classificador possui a melhor performance nesta medida. Estes resultados podem ser verificados na Tabela 5.34 e na Figura 5.17.

Tabela 5.34: Valores estatísticos do método de *Bonferroni* para a precisão dos classificadores.

Classificador	L. Inferior	Média	L. Superior	P-Value
RNA(1) RL(2)	0,285	1,667	3,049	0,012
RNA(1) SVM(3)	-1,049	0,333	1,715	1,000
RL(2) SVM(3)	-2,715	-1,333	0,049	0,063

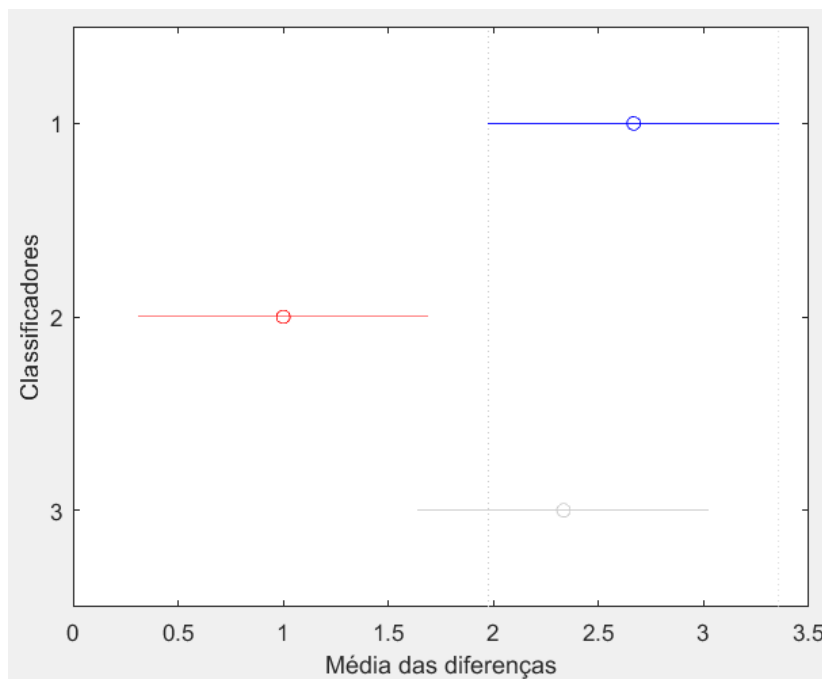


Figura 5.17: Gráfico do método de *Bonferroni* para a precisão dos classificadores.

Pela análise comparativa foi possível verificar que a RNA e o SVM tiveram a mesma performance em termos estatísticos para a precisão, já a RL obteve um desempenho mais baixo.

Na Tabela 5.35 é exposto o resultado do teste de *Friedman* para a acurácia dos classificadores.

Tabela 5.35: Valores estatísticos do teste de *Friedman* para a acurácia dos classificadores.

Nº de Algoritmos=3	χ^2	p-value
Nº de domínios=6		
Teste	10,17	0,006
H0 para 0,5	7	

O valor do χ^2 para a acurácia foi de 10,17, superior ao valor crítico 7, o que significa que existe diferença estatística para a acurácia. Então a análise comparativa com *Bonferroni* foi executada para verificar qual classificador possui a melhor performance nesta medida. Estes resultados podem ser verificados na Tabela 5.36 e na Figura 5.18.

Tabela 5.36: Valores estatísticos do método de *Bonferroni* para a acurácia dos classificadores.

Classificador	L. Inferior	Média	L. Superior	P-Value
RNA(1) RL(2)	0,397	1,750	3,103	0,006
RNA(1) SVM(3)	-0,853	0,500	1,853	1,000
RL(2) SVM(3)	-2,603	-1,250	0,103	0,081

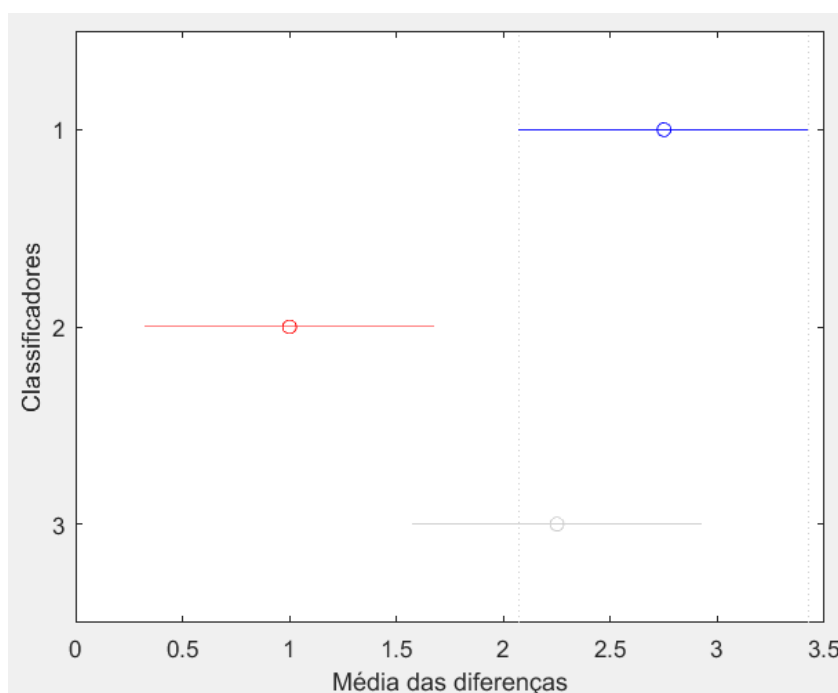


Figura 5.18: Gráfico do método de *Bonferroni* para a acurácia dos classificadores.

Pela análise comparativa foi possível verificar que a RNA e o SVM tiveram a mesma performance em termos estatísticos para a precisão, já a RL obteve um desempenho mais baixo.

Por fim será verificada a performance dos classificadores para a AUC. Primeiro é exibido na Tabela 5.37 o resultado do teste de *Friedman*.

Tabela 5.37: Valores estatísticos do teste de *Friedman* para a AUC dos classificadores.

Nº de Algoritmos=3	χ^2	p-value
Nº de domínios=6		
Teste	10,33	0,005
H0 para 0,5	7	

Neste teste o χ^2 para a acurácia foi de 10,33, sendo, portanto, superior ao valor crítico 7, o que significa que existe diferença estatística para a AUC. Então a análise *post hoc* por *Bonferroni* foi executada para verificar qual classificador possui a melhor performance nesta medida. Estes resultados podem ser verificados na Tabela 5.38 e na Figura 5.19.

Tabela 5.38: Valores estatísticos do método de *Bonferroni* para a AUC dos classificadores.

Classificador	L. Inferior	Média	L. Superior	<i>P-Value</i>
RNA(1) RL(2)	-0,715	0,667	2,049	0,745
RNA(1) SVM(3)	0,451	1,833	3,215	0,004
RL(2) SVM(3)	-0,215	1,167	2,549	0,130

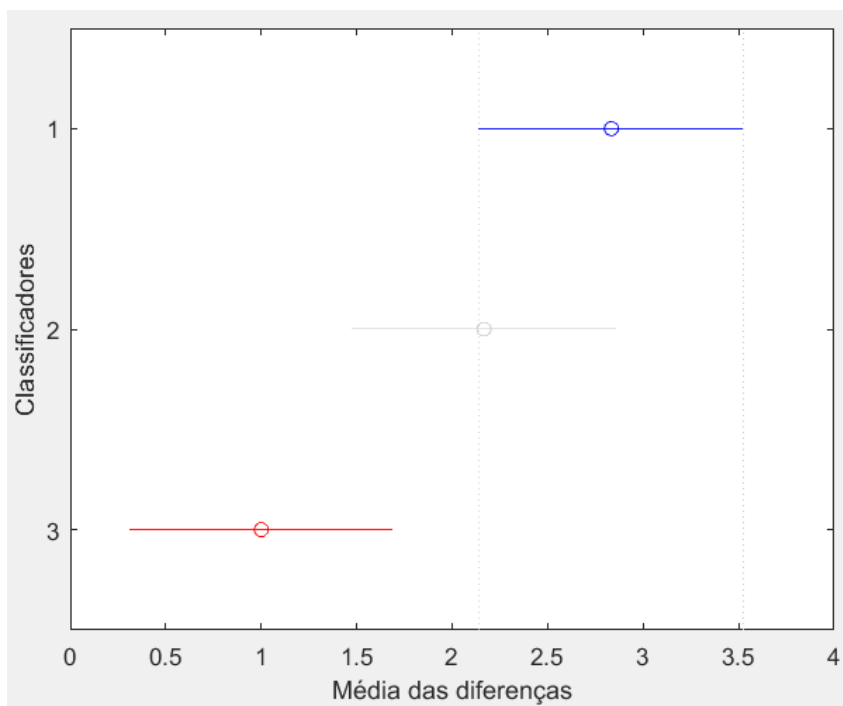


Figura 5.19: Gráfico do método de *Bonferroni* para a AUC dos classificadores.

Nesta análise para a AUC foi observado que a RNA e a RL tiveram resultados estatisticamente semelhantes, porém o SVM teve a menor performance nesta medida.

Por fim, após os testes e análises foi verificado que os resultados da RNA foram consistentes em todas as medidas e no geral apresentaram uma performance melhor que os outros classificadores.

5.4 Comparação com o ALFA

Após as comparações das classificações obtidas pelos algoritmos de aprendizado de máquina aqui testados, a RNA que identificou corretamente aproximadamente 93,0% das linhas em emissão teve seu resultado confrontado com a identificação feita pelo ALFA (WESSON, 2016) sobre a amostra de testes T3. Nesta, o ALFA identificou corretamente apenas 15 linhas das 47, ou seja 31,9% do total. O ALFA foi a abordagem para identificação automática de linhas em emissão mais próxima do proposto nesta pesquisa e o método aqui apresentado sobrepujou em muito esta rotina computacional.

Capítulo 6

AILINE

“A filosofia, quando merece esse nome, não é senão Razão, aperfeiçoada pelo Estudo, pela Aprendizagem e pelo uso das coisas.”

-- Robert Boyle

Após a realização dos testes, uma rotina computacional foi desenvolvida no MATLAB, tendo por nome a sigla AILINE (Algoritmo Inteligente para Identificação de Linhas Espectrais), esta recebe como entrada o espectro em forma de um vetor bidimensional (1^a coluna = comprimento de onda, 2^a coluna = intensidade) e após a execução informa a quantidade de linhas em emissão encontradas, depois gera um gráfico para visualização das linhas identificadas no espectro e cria uma tabela constando o índice do ponto central da linha no espectro, o comprimento de onda da linha observado, a *Full Width Half Maximum* (FWHM) e o fluxo da linha, a FWHM e o fluxo da gaussiana ajustada a linha. É importante observar que as medidas realizadas são primeiro feitas diretamente nas linhas encontradas e depois em gaussianas ajustadas as mesmas, fornecendo, portanto, informações sobre a situação real e a desejada, a Figura 6.1 exibe um exemplo do ajuste gaussiano realizado. A geração dos gráficos é opcional.

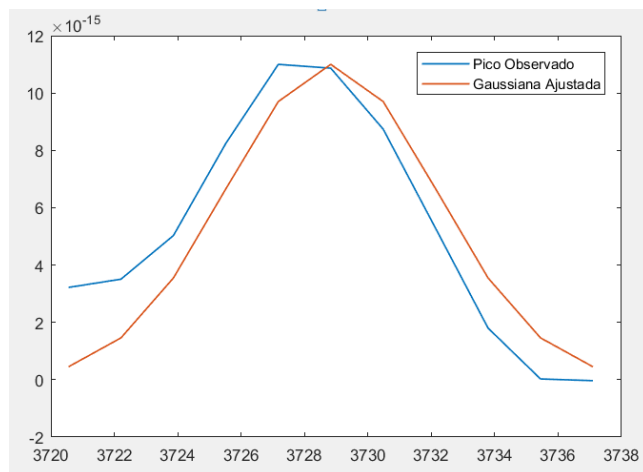


Figura 6.1: Exemplo de gaussiana ajustada em um pico.

Para realizar as medidas, o contínuo é ajustado com uma aproximação *spline* com uma janela de deslocamento com 21 pontos de largura e com passo igual a 21. Estes valores podem ser modificados caso o usuário deseje, a geração do gráfico exibindo o ajuste é opcional. A Figura 6.2 exibe um exemplo de um espectro com o contínuo ajustado.

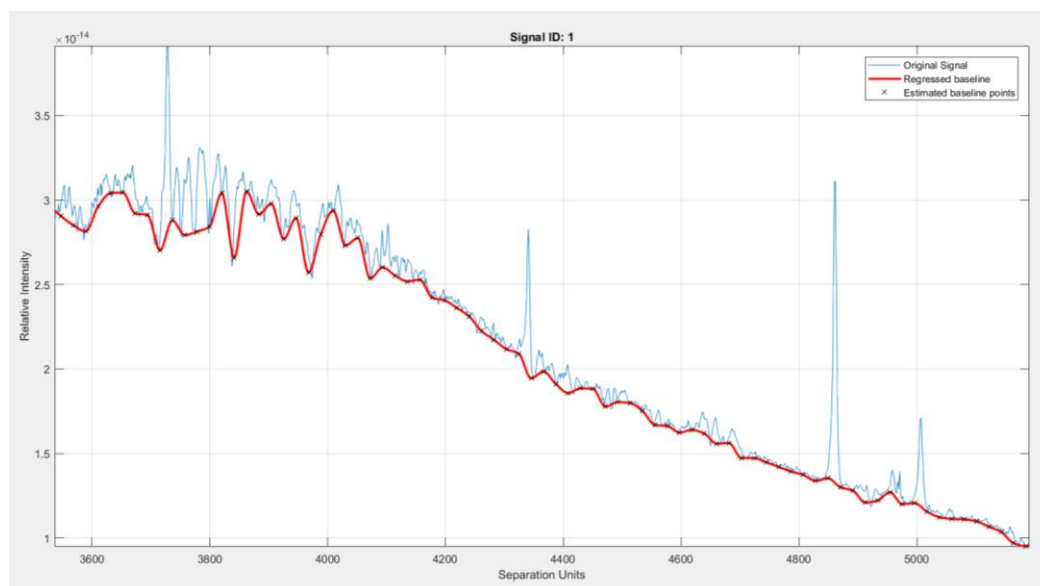


Figura 6.2: Exemplo de ajuste do contínuo.

Se o espectro já estiver calibrado, em comprimento de onda, é possível já associar o íon correspondente com as linhas identificadas com o uso de um catálogo já embutido na rotina conforme mostrado na Figura 6.3 e na Tabela 6.1.

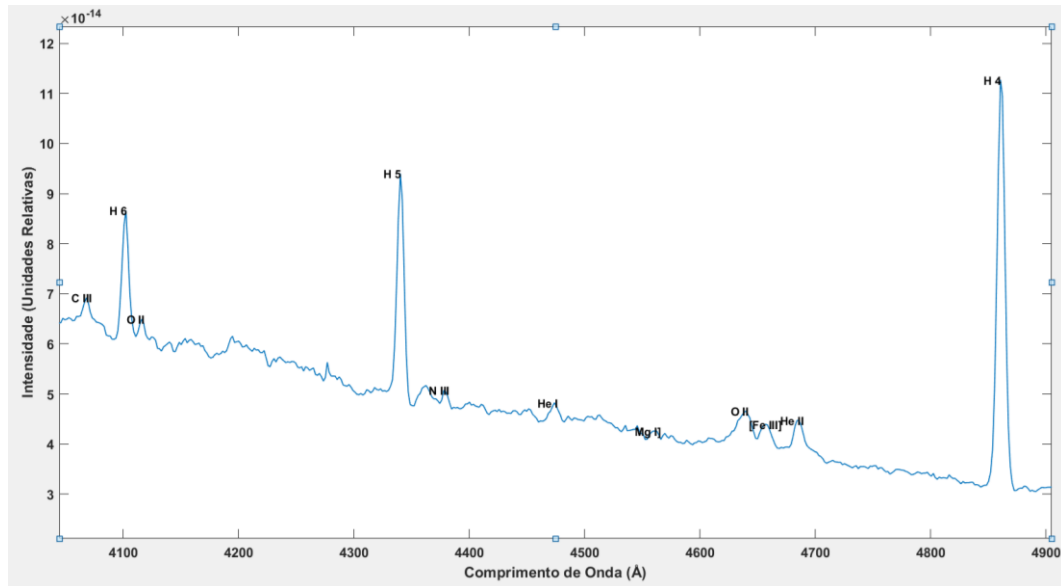


Figura 6.3: Exemplo do gráfico gerado pelo AILINE com identificação dos íons.

Tabela 6.1: Exemplo da tabela gerada pelo AILINE com identificação dos íons.

Índice	Observado	Teórico	Íon	FWHM	Fluxo	FWHMG	FluxoG
243	4068,25	4067,94	C III	7,36302	2,83E-15	6,94741	3,42E-14
263	4101,84	4101,74	H 6	6,37556	2,48E-14	8,98613	2,26E-13
271	4117,44	4119,22	O II	5,72961	2,80E-15	5,90192	2,50E-14
402	4340,13	4340,47	H 5	6,72629	4,36E-14	9,53005	4,32E-13
424	4378,77	4379,11	N III	5,76002	2,66E-15	4,84867	1,76E-14
480	4473,14	4471,50	He I	8,10379	3,63E-15	8,66652	3,21E-14
530	4559,98	4562,60	[Mg I]	8,41682	1,58E-15	5,57877	8,11E-15
577	4637,31	4638,86	O II	12,90469	5,50E-15	13,49345	7,48E-14
587	4656,90	4658,10	[Fe III]	9,17915	3,20E-15	7,31011	2,20E-14
603	4686,49	4685,68	He II	9,27256	5,28E-15	12,91851	8,33E-14
706	4860,78	4861,33	H 4	7,69435	7,95E-14	11,19580	9,21E-13
746	4930,31	4931,80	[O III]	8,64465	1,32E-15	6,68431	8,01E-15
764	4958,06	4958,91	[O III]	8,56740	1,19E-14	11,30829	1,46E-13
773	4974,99	4972,47	[FeVI]	5,25748	7,47E-16	4,51415	3,07E-15
791	5006,22	5006,84	[O III]	8,04827	3,57E-14	15,33041	5,78E-13
903	5198,12	5199,84	[N I]	8,72529	5,04E-15	11,82859	5,99E-14
929	5241,13	0,00	*	10,75207	6,23E-16	6,21464	2,81E-15

Quando o valor da linha reconhecida não estiver no catálogo esta linha será identificada como desconhecida apresentando um “*” no local do Íon da tabela e o valor 0 para o valor teórico conforme exibe a Tabela 6.1.

6.1 Tempo de Execução

Foi realizada a medida da velocidade com que o AILINE processa os resultados para os quatro espectros do conjunto de testes T2. A rotina foi executada 100 vezes para cada espectro em um total de 400 execuções com a geração dos gráficos desabilitada, durando apenas 14,28 segundos. Em uma média de $0,14 \pm 0,02$ segundos por execução. Com a geração dos gráficos este tempo fica em $0,21 \pm 0,02$ segundos por execução. O computador utilizado para este teste foi um notebook com processador Intel Core i7-5500U com 2,40 GHz com dois núcleos e 16 GB de RAM.

6.2 AILINE-*Training*

O AILINE-*Training* é uma rotina computacional adicional que possibilita o treino de uma RNA pelo próprio usuário. Assim, caso os usuários queiram utilizar o AILINE em uma faixa do espectro não abordada nesta pesquisa, ou com espectros que possuam características diferentes das testadas em que talvez os resultados não sejam satisfatórios com a RNA atual, eles podem treinar uma RNA baseada em sua própria base de dados.

Para isto, será necessário apenas um conjunto de espectros com as características desejadas, em forma de vetores bidimensionais, este conjunto deve ser subdividido em dois grupos, um grupo com aproximadamente 80 a 90% da base para ser o grupo de treino e o restante para ser utilizado no grupo de teste. Além dos espectros, deve ser fornecido um catálogo com as linhas em emissão existentes em cada espectro, este catálogo são vetores de uma coluna contendo o comprimento de onda das linhas previamente identificadas.

Fornecendo apenas estas entradas, o AILINE-*Training* retornará a RNA treinada e com as devidas medidas de *Recall*, Precisão, Acurácia e AUC para a

amostra de teste. O treino será realizado com os mesmos parâmetros utilizados e avaliados nesta pesquisa. Esta nova rede pode então ser fornecida para o AILINE em lugar da RNA padrão.

Capítulo 7

Considerações Finais

“Se fui capaz de ver mais longe, é porque me apoiei em ombros de gigantes”

-- Isaac Newton

Com o volume de dados atual é altamente desejável uma abordagem rápida para a identificação das linhas, e como foi exposto, a rotina computacional apresentada nesta pesquisa consegue fazer isto em uma fração de segundos. Em uma abordagem manual ou semiautomática este tempo pode chegar de horas a dias, logo, o uso de uma rotina automática como a apresentada nesta pesquisa auxiliará o pesquisador dando-lhe a possibilidade de usar seu tempo na análise dos dados e não na identificação das linhas.

Além do fator velocidade, a abordagem conseguiu localizar mais de 93% das linhas existentes nos espectros testados, obtendo, nesta amostra, resultados melhores nas medidas de *recall*, precisão, acurácia e AUC do que outros algoritmos de aprendizado de máquina e identificando quase três vezes mais linhas em emissão do que o ALFA, mesmo em espectros em condições diferentes de sinal/ruído e resolução espectral.

Nesta pesquisa foi possível verificar através da revisão da literatura a importância da espectroscopia óptica para a Astronomia, como também foi visto que ao mesmo tempo em que o atual avanço tecnológico permite a coleta de dados a

níveis nunca vistos para esta área, um problema surge, pois se torna humanamente impossível a análise manual destes dados em tempo hábil.

Em complemento, foi destacado que a identificação das linhas em emissão é uma das fases cruciais para a análise dos dados espectroscópicos, porém este processo por meio de rotinas manuais ou semiautomáticas é lento e dependente do grau de experiência do especialista.

Ao serem verificadas as soluções existentes, foi percebido que as mesmas ainda não atendem a demanda, porém foi observado que existe uma crescente tendência no uso de algoritmos de aprendizado de máquina para aplicações diversas na espectroscopia e em outras áreas afins, o que indicou um possível caminho para a identificação automática das linhas em emissão nos espectros ópticos. Este caminho foi perseguido nesta pesquisa por meio da utilização de redes neurais artificiais para classificação de picos como linhas em emissão em espectros ópticos de galáxias, o que se mostrou uma alternativa viável, alcançando bons resultados nas amostras utilizadas nos testes. Estes resultados ao serem avaliados e comparados com outras abordagens se mostraram consistentes e viáveis.

Então, uma rotina computacional foi desenvolvida no Matlab, o AILINE (Algoritmo Inteligente para Identificação de Linhas Espectrais), que recebe como entrada um espectro na forma de um vetor bidimensional e devolve uma tabela com as linhas em emissão identificadas e um gráfico para a visualização dos resultados. Este processo é automático e não necessita de parametrização, tornando-o independente da experiência do usuário. Tudo isto é feito em uma fração de segundos e, como demonstrado nos testes, com uma acurácia acima de 95%.

7.1 Aplicações

O AILINE em sua forma atual já pode ser utilizado para identificação das linhas em emissão em espectros ópticos de galáxias onde o processo de redução já foi realizado até a correção do *redshift*. Como ele provê uma tabela com as linhas identificadas e com algumas medidas básicas, é possível utilizar estas informações como dados iniciais para estimar abundância química, fazer algumas classificações

sobre o tipo de galáxia que se está verificando, estimativas de temperatura e idade e uma ampla gama de análises.

Caso o *redshift* ainda não esteja corrigido, é possível utilizar a informação das linhas encontradas como entrada para outras rotinas que realizam a correção do mesmo.

Esta pesquisa manteve seu escopo nos espectros de galáxias na faixa do óptico, mas como a espectroscopia não é feita apenas na faixa do óptico e nem é uma área exclusiva da Astronomia, este método pode ser estendido para outras faixas do espectro eletromagnético e mesmo para outras áreas em que a análise das linhas espectrais é utilizada, como por exemplo a química, a biomedicina e engenharias diversas. Esta extensão pode ser feita treinando uma RNA utilizando o AILINE-Training, ou outra ferramenta, e fornecendo esta nova RNA para o AILINE, adicionalmente um novo catálogo pode ser fornecido de acordo a faixa do espectro desejada.

7.2 Pesquisas Futuras

A pesquisa aqui desenvolvida demonstrou a viabilidade das redes neurais artificiais para a classificação das linhas em emissão. Para aumentar a confiabilidade nos resultados será interessante a criação de uma base maior e mais diversificada de espectros para uso no treino e testes, esta base poderá servir não apenas para melhorias desta proposta, quanto para o treinamento de outras abordagens que utilizem algoritmos de aprendizado de máquina. Futuramente, pesquisas adicionais podem prover um *upgrade* na abordagem aqui desenvolvida para que as linhas em absorção sejam identificadas de forma semelhante. Ainda é possível complementar a abordagem para que a correção do *redshift* seja também realizado pela mesma. Aliando-se a isto, outras medições das linhas podem ser incluídas tornando a abordagem um *pipeline* completo para análise das linhas espectrais não apenas em galáxias, mas também para estrelas e outros objetos astronômicos. Podendo ainda ser adaptada para ser utilizada nas demais áreas afins da espectroscopia além da Astronomia.

Referências Bibliográficas

- ABDEL-AAL, Radwan E. Comparison of algorithmic and machine learning approaches for the automatic fitting of Gaussian peaks. **Neural computing & applications**, v. 11, n. 1, p. 17-29, 2002.
- ALAM, S. et al. The eleventh and twelfth data releases of the Sloan Digital Sky Survey: final data from SDSS-III. **The Astrophysical Journal Supplement Series**, v. 219, n. 1, p. 12, 2015.
- BORNE, K. D. Astrominformatics: a 21st century approach to astronomy. **arXiv preprint arXiv:0909.3892**, 2009.
- BROMOVÁ, P.; ŠKODA, P.; VÁŽNÝ, J. Classification of spectra of emission line stars using machine learning techniques. **International Journal of Automation and Computing**, v. 11, n. 3, p. 265-273, 2014.
- BYKOV, A. D. et al. Application of pattern recognition in molecular spectroscopy: Automatic line search in high-resolution spectra. **Optics and spectroscopy**, v. 96, n. 4, p. 497-502, 2004.
- CRISTIANINI, Nello; SHAWE-TAYLOR, John. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge university press, 2000.
- DA SILVA, Ivan Nunes et al. **Artificial Neural Networks: A Practical Course**. Springer, 2016.
- DALCANTON, J. et al. From cosmic birth to living earths: the future of UVOIR space astronomy. **arXiv preprint arXiv:1507.04779**, 2015.
- DEMUTH, Howard B. et al. **Neural network design**. Martin Hagan, 2014.
- DRAVINS, D. High-fidelity spectroscopy at the highest resolutions. **Astronomische Nachrichten**, v. 331, n. 5, p. 535-540, 2010.
- DREISEITL, Stephan; OHNO-MACHADO, Lucila. Logistic regression and artificial neural network classification models: a methodology review. **Journal of biomedical informatics**, v. 35, n. 5, p. 352-359, 2002.

- DU, P.; KIBBE, W. A.; LIN, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. **Bioinformatics**, v. 22, n. 17, p. 2059-2065, 2006.
- FERNANDES, I. F. et al. Massive star populations in Wolf-Rayet galaxies. **Monthly Notices of the Royal Astronomical Society**, v. 355, n. 3, p. 728-746, 2004.
- FEKLISTOVA, T. et al. Atomic Line Catalogue for Gaseous Nebulae. **Baltic Astronomy**, v. 3, p. 292, 1994.
- FREDRIKSSON, M. J. et al. An automatic peak finding method for LC-MS data using Gaussian second derivative filtering. **Journal of separation science**, v. 32, n. 22, p. 3906-3918, 2009.
- GONÇALVES, RODRIGO MIKOSZ et al. Modelagem preditiva de linha de costa utilizando redes neurais artificiais. **Boletim de Ciências Geodésicas**, v. 16, n. 3, p. 420-444, 2010.
- HETEM, J. Gregório; PEREIRA, V. Jatenco. **Fundamentos de Astronomia: Distribuição de Energia e Linhas Espectrais**. IAG/USP Departamento de Astronomia, 2010. Disponível em: <<http://www.astro.iag.usp.br/~jane/aga215/apostila/cap05.pdf>> Acesso em: 16 Jul. 2017.
- HONG, S.; DEY, A.; PRESCOTT, M. KM. On the Automated and Objective Detection of Emission Lines in Faint-Object Spectroscopy. **Publications of the Astronomical Society of the Pacific**, v. 126, n. 945, p. 1048, 2014.
- HOPKINS, A. M. et al. Galaxy And Mass Assembly (GAMA): spectroscopic analysis. **Monthly Notices of the Royal Astronomical Society**, v. 430, n. 3, p. 2047-2066, 2013.
- JAPKOWICZ, Nathalie; SHAH, Mohak. **Evaluating learning algorithms: a classification perspective**. Cambridge University Press, 2011.
- KITCHIN, C. R. **Optical astronomical spectroscopy**. Philadelphia: IOP Publishing Ltd, 1995.
- LIVINGSTONE, David J. **Artificial Neural Networks: Methods and Applications (Methods in Molecular Biology)**. Humana Press, 2008.
- LUO, A. L. et al. The first data release (DR1) of the LAMOST regular survey. **Research in Astronomy and Astrophysics**, v. 15, n. 8, p. 1095, 2015.
- MAIRE, J. et al. Data reduction pipeline for the Gemini Planet Imager. In: **SPIE Astronomical Telescopes Instrumentation**. International Society for Optics

- and Photonics, 2010. p. 773531-773531-11.
- MARGOTTO, P. R. Boletim Informativo Pediátrico. **Curva ROC: Como Fazer e Interpretar no SPSS.** 2009. Disponível em: http://www.paulomargotto.com.br/documentos/BIP_72-ANO_29-2009.pdf. Acesso em 23/07/2017.
- MTETWA, N.; SMITH, L. S. Smoothing and thresholding in neuronal spike detection. **Neurocomputing**, v. 69, n. 10, p. 1366-1370, 2006.
- PANOULAS, K. I.; HADJILEONTIADIS, L. J.; PANAS, S. M. Enhancement of R-wave detection in ECG data analysis using higher-order statistics. In: **Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE.** IEEE, 2001. p. 344-347.
- PLANO NACIONAL DE ASTRONOMIA. **Proposta da Comissão Especial de Astronomia.** Presidência da República; Ministério da Ciência e Tecnologia, 2010. Disponível em: <<http://www.lna.br/PNA-FINAL.pdf>> Acesso em: 01 fev. 2016.
- REQUENA, A.; ZÚÑIGA, J. **Química Física. Problemas de espectroscopia.** Madri: Pearson Educación S.A, 2007.
- SALA, O. **Fundamentos da espectroscopia Raman e no infravermelho.** São Paulo: UNESP, 2008.
- SCHOLKMANN, F.; BOSS, J.; WOLF, M. An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. **Algorithms**, v. 5, n. 4, p. 588-603, 2012.
- SHARPEE, B. et al. Introducing EMILI: Computer-aided Emission Line Identification. **The Astrophysical Journal Supplement Series**, v. 149, n. 1, p. 157, 2003.
- SHIM, B.; MIN, H.; YOON, S. Nonlinear preprocessing method for detecting peaks from gas chromatograms. **BMC bioinformatics**, v. 10, n. 1, p. 1, 2009.
- ŠKODA, P.; VÁŽNÝ, J. Searching of new emission-line stars using the astroinformatics approach. **arXiv preprint arXiv:1112.2775**, 2011.
- STASINSKA, G. What can emission lines tell us?. **arXiv preprint arXiv:0704.0348**, 2007.
- TU, Liangping et al. Automatic Classification of Stellar Spectra used Neural Network. In: 2008 **Fourth International Conference on Natural Computation.** IEEE, 2008. p. 105-109.

- VAN OUYEN, Arjen; NIENHUIS, Bernard. Improving the convergence of the back-propagation algorithm. **Neural Networks**, v. 5, n. 3, p. 465-471, 1992.
- VIJAYA, G.; KUMAR, V.; VERMA, H. K. ANN-based QRS-complex analysis of ECG. **Journal of medical engineering & technology**, v. 22, n. 4, p. 160-167, 1998.
- VIVÓ-TRUYOLS, G. et al. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection. **Journal of Chromatography A**, v. 1096, n. 1, p. 133-145, 2005.
- VON BERLEPSCH, R. (Ed.). **Reviews in Modern Astronomy, Deciphering the Universe through Spectroscopy**. John Wiley & Sons, 2011.
- YORK, D. G. et al. The sloan digital sky survey: Technical summary. **The Astronomical Journal**, v. 120, n. 3, p. 1579, 2000.
- WESSON, R. Alfa: an automated line fitting algorithm. **Monthly Notices of the Royal Astronomical Society**, v. 456, n. 4, p. 3774-3781, 2016.
- WUNSCH, Stefan; FINK, Johannes; JONDRAL, Friedrich K. Improved Detection by Peak Shape Recognition Using Artificial Neural Networks. In: **Vehicular Technology Conference (VTC Fall)**, 2015 IEEE 82nd. IEEE, 2015. p. 1-5.
- ZHANG, Jifu et al. A concept lattice based outlier mining method in low-dimensional subspaces. **Pattern Recognition Letters**, v. 30, n. 15, p. 1434-1439, 2009.
- ZHAO, G. et al. Stellar abundance and Galactic chemical evolution through LAMOST spectroscopic survey. **Chinese Journal of Astronomy and Astrophysics**, v. 6, n. 3, p. 265, 2006.

Apêndice A

Espectros Utilizados na Pesquisa

