



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2019/04.04.01.38-TDI

CLASSIFICAÇÃO INTELIGENTE DE SUPERNOVAS UTILIZANDO HIERARQUIA DE REDES NEURAIAS ARTIFICIAIS

Francisca Joamila Brito do Nascimento

Dissertação de Mestrado do
Curso de Pós-Graduação em
Computação Aplicada, orientada
pelo Dr. Lamartine Nogueira
Frutuoso Guimarães, aprovada em
10 de maio de 2019.

URL do documento original:

<<http://urlib.net/8JMKD3MGP3W34R/3T3PTLP>>

INPE
São José dos Campos
2019

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE

Gabinete do Diretor (GBDIR)

Serviço de Informação e Documentação (SESID)

CEP 12.227-010

São José dos Campos - SP - Brasil

Tel.:(012) 3208-6923/7348

E-mail: pubtc@inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE - CEPPII (PORTARIA Nº 176/2018/SEI-INPE):**Presidente:**

Dra. Marley Cavalcante de Lima Moscati - Centro de Previsão de Tempo e Estudos Climáticos (CGCPT)

Membros:

Dra. Carina Barros Mello - Coordenação de Laboratórios Associados (COCTE)

Dr. Alisson Dal Lago - Coordenação-Geral de Ciências Espaciais e Atmosféricas (CGCEA)

Dr. Evandro Albiach Branco - Centro de Ciência do Sistema Terrestre (COCST)

Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia e Tecnologia Espacial (CGETE)

Dr. Hermann Johann Heinrich Kux - Coordenação-Geral de Observação da Terra (CGOBT)

Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação - (CPG)

Silvia Castro Marcelino - Serviço de Informação e Documentação (SESID)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon

Clayton Martins Pereira - Serviço de Informação e Documentação (SESID)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Simone Angélica Del Ducca Barbedo - Serviço de Informação e Documentação (SESID)

André Luis Dias Fernandes - Serviço de Informação e Documentação (SESID)

EDITORAÇÃO ELETRÔNICA:

Ivone Martins - Serviço de Informação e Documentação (SESID)

Cauê Silva Fróes - Serviço de Informação e Documentação (SESID)



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21c/2019/04.04.01.38-TDI

CLASSIFICAÇÃO INTELIGENTE DE SUPERNOVAS UTILIZANDO HIERARQUIA DE REDES NEURAIAS ARTIFICIAIS

Francisca Joamila Brito do Nascimento

Dissertação de Mestrado do
Curso de Pós-Graduação em
Computação Aplicada, orientada
pelo Dr. Lamartine Nogueira
Frutuoso Guimarães, aprovada em
10 de maio de 2019.

URL do documento original:

<<http://urlib.net/8JMKD3MGP3W34R/3T3PTLP>>

INPE
São José dos Campos
2019

Dados Internacionais de Catalogação na Publicação (CIP)

Nascimento, Francisca Joamila Brito do.

Na17c Classificação inteligente de supernovas utilizando hierarquia de
redes neurais artificiais / Francisca Joamila Brito do Nascimento.
– São José dos Campos : INPE, 2019.
xxiv + 96 p. ; (sid.inpe.br/mtc-m21c/2019/04.04.01.38-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto
Nacional de Pesquisas Espaciais, São José dos Campos, 2019.
Orientador : Dr. Lamartine Nogueira Frutuoso Guimarães.

1. Supernovas. 2. Classificação automática. 3. Inteligência
artificial. 4. Redes neurais artificiais. I.Título.

CDU 004.8:524.352



Esta obra foi licenciada sob uma Licença [Creative Commons Atribuição-NãoComercial 3.0 Não Adaptada](https://creativecommons.org/licenses/by-nc/3.0/).

This work is licensed under a [Creative Commons Attribution-NonCommercial 3.0 Unported License](https://creativecommons.org/licenses/by-nc/3.0/).

Aluno (a): **Francisca Joamila Brito do Nascimento**

Título: "CLASSIFICAÇÃO INTELIGENTE DE SUPERNOVAS UTILIZANDO HIERARQUIA DE REDES NEURAIS ARTIFICIAIS"

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de **Mestre** em **Computação Aplicada**

Dr. Haroldo Fraga de Campos Velho

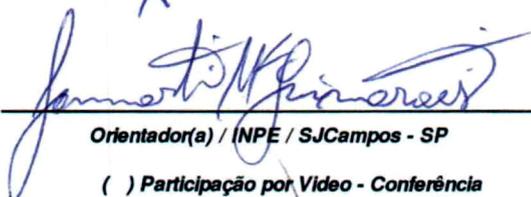


Presidente / INPE / São José dos Campos - SP

() Participação por Vídeo - Conferência

Aprovado () Reprovado

Dr. Lamartine Nogueira Frutuoso Guimarães

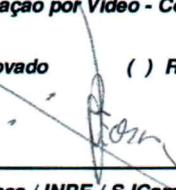


Orientador(a) / INPE / SJC Campos - SP

() Participação por Vídeo - Conferência

Aprovado () Reprovado

Dr. Reinaldo Roberto Rosa



Membro da Banca / INPE / SJC Campos - SP

() Participação por Vídeo - Conferência

Aprovado () Reprovado

Dra. Paula Rodrigues Teixeira Coelho



Convidado(a) / IAG / São Paulo - SP

() Participação por Vídeo - Conferência

Aprovado () Reprovado

Dr. Marcelo Módolo



Convidado(a) / UESP / São Bernardo do Campo - SP

() Participação por Vídeo - Conferência

Aprovado () Reprovado

Este trabalho foi aprovado por:

() maioria simples

unanimidade

São José dos Campos, 10 de maio de 2019

*“No meio do caminho tinha uma pedra”,/ mas a ousada esperança/
de quem marcha cordilheiras/ triturando todas as pedras/ da primeira
à derradeira/ de quem banha a vida toda/ no unguento da coragem/
e da luta cotidiana/ faz do sumo beberagem/ topa a pedra-pesadelo/ é
ali que faz parada/ para o salto e não o recuo/ não estanca os seus
sonhos/ lá no fundo da memória,/ pedra, pau, espinho e grade/ são
da vida desafio./ E se cai, nunca se perdem,/ pois os seus sonhos
esparramados/ adubam a vida, multiplicam/ são motivos de viagem”.*

CONCEIÇÃO EVARISTO

“Pedra, pau, espinho e grade, em: Poemas da recordação e
outros movimentos”, 2008.

*Dedico a minha avó **Alice** (em memória), a minha mãe
Joanalice e a todas as mulheres negras que existiram e
resistiram antes de mim*

AGRADECIMENTOS

Agradeço a meus pais Joanalice e Valmir por me apoiarem em todos os momentos e me receberem amorosamente nas minhas voltas para casa. Agradecimentos aos meus irmãos e cunhadas por me ajudarem sempre que necessário. Agradeço também a minha avó Alice (em memória) por durante a sua vida sempre ter estado por perto.

Aos amigos de Caucaia/Fortaleza (e outras cidades) agradeço por continuarem presentes apesar da distância física. Agradeço aos amigos de São José dos Campos que tornaram a estadia longe de casa mais agradável. Em especial, agradeço ao clube de leitura Leia Mulheres de São José dos Campos que me apresentaram histórias de mulheres fortes que me inspiraram a também ser forte.

Agradeço ao Me. Luis Ricardo Arantes Filho que me ajudou na compreensão de conceitos essenciais da pesquisa e me deu conselhos valiosos sobre como lidar com as venturas e desventuras da pós-graduação. Agradeço aos professores do INPE que sempre foram muito profissionais e dedicados na transmissão de seus conhecimentos.

Agradeço a todos os professores que passaram pela minha vida e contribuíram para a minha formação. Especialmente os que não se preocuparam apenas com conhecimentos técnicos, mas os que entenderam que a formação humana é tão ou mais importante do que a profissional.

Ao INPE e ao DCTA/IEAv, agradeço pelo espaço físico e recursos disponíveis para o desenvolvimento da minha pesquisa. E agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo suporte financeiro dedicado à minha educação.

RESUMO

Uma supernova corresponde à fase final da vida de algumas estrelas, o auge dessa fase é marcado por uma explosão de intenso brilho. O evento da supernova recebe bastante atenção dos estudiosos em Astronomia e Cosmologia, principalmente as supernovas do tipo Ia. A classificação das supernovas as divide em dois tipos principais, as do tipo I não apresentam Hidrogênio no espectro enquanto as do tipo II apresentam. Além da divisão nesses dois tipos, há ainda uma subdivisão que estabelece os tipos Ia, Ib e Ic. Na prática, a classificação das supernovas exige conhecimento especializado de astrônomos e dados (espectros de luz) de boa qualidade. O que se observa nos espectros para fazer a classificação são as linhas de emissão (picos) e absorção (vales) de alguns elementos químicos, como Hidrogênio, Silício, Enxofre e Hélio. Alguns classificadores inteligentes foram desenvolvidos e são reportados na literatura, um deles é a CIntIa (Classificador Inteligente de supernovas do tipo Ia), proposta por (MÉDOLO, 2016). A CIntIa usa redes neurais artificiais binárias para classificar as supernovas nos tipos Ia, Ib, Ic e II com atenção especial para as do tipo Ia. Este trabalho, tem por objetivo aperfeiçoar o sistema CIntIa a fim de que ele seja menos restrito, a sua capacidade de generalização seja expandida e a classificação não produza respostas ambíguas. Para alcançar esse objetivo realizamos algumas etapas, como mudança na variação do comprimento de onda dos espectros aceitos, mudança na estratégia de filtragem e implementação de uma arquitetura hierárquica de redes neurais binárias. Os resultados da classificação dos espectros de supernovas do tipo Ia e II são excelentes. No entanto a classificação de espectros dos tipos Ib e Ic não apresenta bons resultados, o que confirma estudos teóricos que afirmam que os espectros de SNs Ib e Ic não apresentam padrões bem estabelecidos. O aumento expressivo da quantidade de dados (a quantidade de espectros foi incrementada em mais de 1300%, de 649 para 9156) é um fator fundamental para que a análise dos resultados seja mais segura. Outro resultado importante alcançado foi a inclusão de espectros antes e depois do brilho máximo nos treinamentos (a classificação na fase de brilho máximo é o procedimento padrão para todos os classificadores pesquisados). O novo sistema automático e inteligente, originado da CIntIa, chama-se CINTIA 2 e foi implementado nas linguagens de programação C++ e Python, podendo ser utilizado em conjunto com telescópios e espectroscópios diversos.

Palavras-chave: Supernovas. Classificação automática. Inteligência artificial. Redes neurais artificiais.

INTELLIGENT CLASSIFICATION OF SUPERNOVAE USING A HIERARCHY OF ARTIFICIAL NEURAL NETWORKS

ABSTRACT

A supernova corresponds to the final phase of the life of some stars, the peak of this phase is marked by an explosion of intense brightness. The supernova event receives a lot of attention from researchers in Astronomy and Cosmology, mainly type Ia supernovae. The classification of supernovae divides them into two main types, those of type I do not present Hydrogen in the spectrum while those of type II present. In addition to the division into these two types, there is still a subdivision that establishes types Ia, Ib and Ic. In practice, the classification of supernovae requires specializing knowledge of astronomers and data (light spectra) of good quality. What is observed in the spectra to make the classification are the emission lines (peaks) and absorption (valleys) of some chemical elements, such as Hydrogen, Silicon, Sulfur and Helium. Some intelligent classifiers have been developed and are reported in the literature, one of them is CIntIa (Classificador Inteligente de Supernovas do tipo Ia, in Portuguese), proposed by (MÓDOLO, 2016). CIntIa uses binary artificial neural networks to classify supernovas in the types Ia, Ib, Ic and II with special attention to the type Ia. This work aims to improve CIntIa system so that it will have been less restricted, their generalization will have been expanded and classification won't have produced ambiguous answers. In order to achieve this goal, we performed several steps, such as changing the wavelength variation of the accepted spectra, changing the filtering strategy and implementing a hierarchical architecture of binary neural networks. The results of the classification of the type Ia and II supernova spectra are excellent. However, the classification of spectra of types Ib and Ic does not present good results, which confirms theoretical studies that affirm that the spectra of SNs Ib and Ic do not present well-established patterns. The significant increase in the amount of data (the number of spectra was increased by more than 1300%, from 649 to 9156) was a fundamental factor to make the analysis of the results safer. Another important result was the inclusion of spectra before and after the maximum brightness in the training (the classification in the maximum brightness phase is the standard procedure for all classifiers researched). The new intelligent and automatic system, originated from CIntIa, is called CINTIA 2 and was implemented in programming languages C ++ and Python, and can be used in conjunction with telescopes and various spectroscopes.

Keywords: Supernovae. Automatic classification. Artificial intelligence. Artificial neural networks.

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Camadas de uma estrela massiva.	5
2.2 Fluxos de evolução da estrela de acordo com a sua massa.	6
2.3 Esquema de Turatto para classificação de supernovas.	7
2.4 Padrão dos espectros de SNs dos tipos Ia, Ib, Ic, Ic-bl, II e IIb.	9
2.5 Evolução dos espectros das SNs 1994D e 1987L.	9
3.1 Neurônio de McCulloch-Pitts.	12
3.2 Modelo de um neurônio artificial.	13
3.3 Representação de uma RNA com mais de uma camada.	14
3.4 Representação da retropropagação em um MLP com uma camada oculta.	16
3.5 Como calcular a correção dos pesos no algoritmo de retropropagação.	17
3.6 Fase de decomposição de um problema multi-classe que deve classificar bolas nas cores azul, verde ou vermelha.	18
4.1 Resultado da classificação da SN 2002bo pelo GELATO.	22
4.2 Resultado da classificação feita pelo DRACULA (à esquerda) e as sub- classes do tipo Ia propostas por (WANG et al., 2009) (à direita).	25
4.3 Funções de pertencimento e variáveis linguísticas das variáveis de entrada Intensidade de Pico e Largura Equivalente.	27
4.4 Função de pertencimento e variáveis linguísticas da variável de entrada Distância Relativa.	28
5.1 Quantidade de SNs (esquerda) e de espectros (direita) por tipo antes da limpeza, depois da limpeza do conjunto de dados e após aplicar o critério de seleção de intervalo de comprimento de onda.	31
5.2 Histogramas de fase espectral por tipo de SNs.	34
5.3 Resultado da dupla filtragem de um espectro.	35
7.1 Representação gráfica dos intervalos escolhidos como entradas das 4 RNAs. Começando da esquerda os intervalos são referentes às RNAs: Ia, Ib, Ic e II.	42
7.2 Módulos que compõem a CINTIA 2.	54
7.3 Arquitetura da CINTIA 2 desenhada como um fluxograma.	55
8.1 Fluxograma que representa o funcionamento do software CINTIA 2.	58
8.2 Tela que apresenta os resultados da classificação feita pela CINTIA 2.	59
8.3 Visualização dos dados realizada pela CINTIA 2.	61

8.4	Quantidade de dados baixados para a etapa de validação da ferramenta.	62
9.1	Comparativo entre quantidade de SNs (esquerda) e quantidade de espectros (direita) utilizados no desenvolvimento da CIntIa e da CINTIA 2.	65
9.2	Comparação entre os intervalos de comprimento de onda usados, de um espectro da SN1994D, pela CIntIa (acima) e CINTIA 2 (abaixo).	67
9.3	Comparação entre estratégias de filtragem de espectros de SNs, aplicadas em um espectro da SN1998dx.	68
9.4	Comparação entre a filtragem de um espectro da SN1994D feita por Médias Móveis (acima) e Dupla-Filtragem Savitzky-Golay (abaixo).	70
10.1	Espectros da SN1994D em 6 fases espectrais diferentes. Cada fase está em um bloco de fases de acordo com a Tabela 7.2.	76
10.2	Espectros da SN1994D mostrados no mesmo espaço para comparação das mudanças nas linhas espectrais no decorrer do tempo.	77
10.3	Gráfico de dispersão de espectros de SNs na fase de pré-pré-máximo e não-pré-pré-máximo.	80
10.4	Gráfico de dispersão de espectros de SNs na fase de pré-máximo e não-pré-máximo.	81
10.5	Gráfico de dispersão de espectros de SNs na fase de máximo e não-máximo.	81
10.6	Gráfico de dispersão de espectros de SNs na fase de pós-máximo e não-pós-máximo.	82
10.7	Gráfico de dispersão de espectros de SNs na fase de pré-nebular e não-pré-nebular.	82
10.8	Gráfico de dispersão de espectros de SNs na fase de nebular e não-nebular.	83
10.9	Gráfico de dispersão 3D dos espectros separados em seis fases espectrais, cada dimensão é um componente principal obtida por meio do algoritmo Isomap.	85

LISTA DE TABELAS

	<u>Pág.</u>
3.1 Funções de ativação mais usadas.	14
4.1 Avaliação das 4 RNAs que compõem o sistema CIntIa.	24
5.1 Bancos de espectros compilados no TOSC e suas principais referências.	29
6.1 Modelo de uma matriz de confusão.	37
6.2 Interpretação do índice Kappa.	38
7.1 Número de neurônios em cada topologia testada.	43
7.2 Bloco de fases espectrais definidas para auxiliar na diversificação dos espectros de SNIa treinados.	44
7.3 Melhores resultados das sessões de treinamento realizadas para definir a amplitude das fases espectrais dos espectros do tipo Ia incluídos no classificador.	44
7.4 Matriz de confusão da RNA Ia adaptada para os 2 testes.	45
7.5 Métricas de avaliação do desempenho da RNA Ia dos 2 testes.	46
7.6 Matriz de confusão da RNA Ib com espectros Ib, apenas na fase de brilho máximo, adaptada para os 2 testes.	46
7.7 Métricas de avaliação do desempenho da RNA Ib, apenas na fase de brilho máximo, dos 2 testes.	47
7.8 Matriz de confusão da RNA Ib adaptada para os 2 testes.	47
7.9 Métricas de avaliação do desempenho da RNA Ib dos 2 testes.	47
7.10 Matriz de confusão da RNA Ic com espectros Ic, apenas na fase de brilho máximo, adaptada para os 2 testes.	48
7.11 Métricas de avaliação do desempenho da RNA Ic, apenas na fase de brilho máximo, dos 2 testes.	49
7.12 Matriz de confusão da RNA Ic adaptada para os 2 testes.	49
7.13 Métricas de avaliação do desempenho da RNA Ic dos 2 testes.	50
7.14 Matriz de confusão da RNA II com espectros II, apenas na fase de brilho máximo, adaptada para os 2 testes.	51
7.15 Métricas de avaliação do desempenho da RNA II, apenas na fase de brilho máximo, dos 2 testes.	51
7.16 Matriz de confusão da RNA II adaptada para os 2 testes.	52
7.17 Métricas de avaliação do desempenho da RNA II dos 2 testes.	52
7.18 Métricas de avaliação do desempenho da CINTIA 2.	53

8.1	Matriz de confusão do teste realizado para validar a ferramenta.	62
8.2	Métricas de avaliação do desempenho da CINTIA 2 ao ser submetida à validação.	63
9.1	Comparação entre quantidade de espectros, fases espectrais e principais métricas de avaliação da RNA Ia da CIntIa; o treino chamado de Intermediário Ia, cujos espectros do tipo Ia estão na fases > -4 e $< +4$ dias; e o módulo Ia da CINTIA 2.	68
9.2	Comparação entre métricas de avaliação dos métodos de filtragem Dupla-Filtragem e Médias Móveis aplicados a espectros de SNs.	69
9.3	Identificadores dos classificadores automáticos para utilização na Tabela 9.4.	71
9.4	Comparação entre CINTIA 2 e outros classificadores automáticos de SNs. "NA"significa não aplica e "NC", não consta. "IA"significa Inteligência Artificial.	71
9.5	Comparação entre as matrizes de confusão de um experimento que aplica votação distribuída na fase de reconstrução da estratégia Um Contra Todos e outro que aplica a hierarquia de RNAs definida neste trabalho. .	73
10.1	Intervalos do comprimento de onda usadas como SFA para calcular fases espectrais.	75
10.2	Quantidade de espectros utilizados na classificação das fases espectrais de SNs Ia.	78
10.3	Matrizes de confusão dos testes das seis RNAs binárias que classificam as fases de espectros de SNs Ia.	78
10.4	Métricas de avaliação do desempenho das seis RNAs binárias que classificam as fases espectrais de SNs Ia. Os mnemônicos que designam as RNAs tem as seguintes significações. PrePreMax: Pré-Pré-Máximo; PreMax: Pré-Máximo; Max: Máximo; PosMax: Pós-Máximo; PreNeb: Pré-Nebular; Neb: Nebular.	79
10.5	Matrizes de confusão das classificações binárias realizadas pelo algoritmo K-Means.	83
10.6	Métricas de avaliação do desempenho dos seis classificadores binários implementados com K-Means.	84

LISTA DE ABREVIATURAS E SIGLAS

CIntIa	–	Classificador Inteligente de supernovas do tipo Ia
SUZAN	–	Sistema fUZzy Avaliador de superNovas
SN	–	Supernova
SNIa	–	Supernova do tipo Ia
SNIb	–	Supernova do tipo Ib
SNIc	–	Supernova do tipo Ic
SNII	–	Supernova do tipo II
H	–	Hidrogênio
He	–	Hélio
C	–	Carbono
O	–	Oxigênio
Ne	–	Neônio
Mg	–	Magnésio
Si	–	Silício
Fe	–	Ferro
Ca	–	Cálcio
Co	–	Cobalto
KDUST	–	Kulun Dark Universe Survey Telescope
RNA	–	Rede Neural Artificial
MLP	–	Multi-Layer Perceptron
SNID	–	Supernova Identification
GELATO	–	GENeric cLAssification TOol
DRACULA	–	Dimensionality Reduction and Clustering for Unsupervised Learning in Astronomy
DL	–	Deep Learning
Cfa	–	Harvard-Smithsonian Center for Astrophysics
AUC	–	Area Under the Curve
TOSC	–	The Open Supernova Catalog
CCCP	–	Caltech Core-Collapse Program
CPCS	–	Cambridge Photometry Calibration Server
CSP	–	Carnegie Supernova Project
SNF	–	Nearby Supernova Factory
OGLE-IV	–	Optical Gravitational Lensing Experiment IV
Pan-STARRS	–	Panoramic Survey Telescope & Rapid Response System
SDSS	–	Sloan Digital Sky Survey
SNHunt	–	Supernova Hunt
SNLS	–	Supernova Legacy Survey
SUSPECT	–	The Online Supernova Spectrum Archive
SNDB	–	UC Berkeley Filippenko Group's Supernova Database
WISeREP	–	Weizmann Interactive Supernova data REPOSITORY

LISTA DE SÍMBOLOS

- \odot - massa solar
- \AA - Angstrom
- λ - comprimento de onda
- μ - micro
- M - macro

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	1
2 SUPERNOVAS	5
2.1 Classificação das Supernovas	6
3 REDES NEURAIAS ARTIFICIAIS	11
3.1 Histórico	11
3.2 Conceitos de Redes Neurais Artificiais	12
3.3 Perceptron de Múltiplas Camadas (MLP)	15
3.4 Estratégia Um Contra Todos para Problemas Multi-classe	17
3.4.1 Fase de Decomposição	17
3.4.2 Fase de Reconstrução	18
4 CLASSIFICADORES AUTOMÁTICOS DE SUPERNOVAS	21
4.1 Supernova Identification (SNID)	21
4.2 Generic Classification Tool (GELATO)	22
4.3 Classificador Inteligente de supernovas do tipo Ia (CIntIa)	23
4.4 Dimensionality Reduction and Clustering for Unsupervised Learning in Astronomy (DRACULA)	24
4.5 Quantitative Classification of Type I Supernovae (Quantitative)	25
4.6 Sistema Fuzzy Avaliador de Supernovas (SUZAN)	26
5 DADOS	29
5.1 Limpeza do Conjunto de Dados	30
5.2 Seleção de Intervalo de Comprimento de Onda	30
5.3 Pré-Processamento	31
5.3.1 Ajuste do Redshift	32
5.3.2 Dupla-Filtragem com Filtro Savitzky-Golay	32
6 MÉTRICAS DE AVALIAÇÃO DO DESEMPENHO	37
7 DESENVOLVIMENTO DA CINTIA 2	41
7.1 Extração das Entradas	41
7.2 Parâmetros de Treinamento	42

7.3	Topologias Testadas	43
7.4	Treinamento e Teste da RNA Ia	43
7.4.1	Detalhamento do Melhor Resultado	45
7.5	Treinamento e Teste da RNA Ib	45
7.6	Treinamento e Teste da RNA Ic	48
7.7	Treinamento e Teste da RNA II	50
7.8	Arquitetura da CINTIA 2	52
8	DESENVOLVIMENTO DO SOFTWARE CINTIA 2	57
8.1	Entrada e Saída	58
8.2	Processamento	59
8.3	Visualização dos Dados	60
8.4	Validação do Software	61
9	DISCUSSÃO SOBRE OS RESULTADOS	65
9.1	CINTIA 2 <i>versus</i> CIntIa	65
9.1.1	Mudança no Intervalo de Comprimento de Onda	66
9.1.2	Aumento da Amplitude das Fases Espectrais	66
9.1.3	Mudança no Método de Filtragem dos Espectros	68
9.2	CINTIA 2 <i>versus</i> Outros Classificadores	70
9.3	Considerações sobre a Estratégia Um Contra Todos	72
10	CONSIDERAÇÕES SOBRE A CLASSIFICAÇÃO DE FASES ESPECTRAIS DE SUPERNOVAS Ia	75
10.1	Classificação com RNAs Binárias	77
10.2	Classificação com K-Means	79
10.3	Dificuldades de Classificação das Fases Espectrais	84
11	CONCLUSÃO	87
	REFERÊNCIAS BIBLIOGRÁFICAS	91

1 INTRODUÇÃO

As supernovas (SNs) são grandes explosões que caracterizam o fim da vida de estrelas muito massivas, em sua maioria. Esses eventos são observados desde épocas remotas, o astrônomo dinamarquês Tycho Brahe (1546-1601) foi um dos observadores e por causa dele a supernova (SN) 1572, que aconteceu na constelação de Cassiopeia, também é conhecida como SN de Tycho Brahe.

A palavra supernova foi criada muitos anos depois do fenômeno observado por Brahe e seus contemporâneos, os autores da nomenclatura são Walter Baade (1893-1960) e Fritz Zwicky (1898-1974). A escolha de “nova” foi baseada na crença anterior de que as SNs eram novas estrelas temporárias e o prefixo “super” foi adicionado por causa do brilho mais intenso do que as “novas” comuns. O brilho intenso das SNs tem uma grande importância nos estudos de Cosmologia. O uso de SNs, por exemplo, possibilitou a descoberta da expansão acelerada do Universo que rendeu o Prêmio Nobel de Física de 2011 para os astrônomos Saul Perlmutter ([PERLMUTTER et al., 1999](#)), Adam G. Riess e Brian P. Schmidt ([RIESS et al., 1998](#)).

As SNs usadas na descoberta da expansão acelerada do Universo foram as do tipo Ia (SNIa). A curva de luz de todas as SNs Ia estudadas são semelhantes, essa característica permite que sejam utilizadas como vela-padrão, um objeto astronômico que por sua luminosidade conhecida é usado para medir distâncias astronômicas. Segundo ([OLIVEIRA FILHO; SARAIVA, 2014](#)), as SNs são classificadas em dois tipos principais propostos por Rudolph Leo Bernhard Minkowski (1895-1976) em 1941, as SNs do tipo I não possuem Hidrogênio (H) e as do tipo II possuem este elemento. Há ainda uma sub-classificação do tipo I que resulta nos tipos Ia, Ib e Ic.

Atualmente, a classificação de SNs segue um consenso de separar entre as de origem termonuclear (Ia) e as outras, originadas do colapso do núcleo. ([TURATTO, 2003](#)) apresenta um esquema de classificação que é seguido por grande parte dos classificadores automáticos e inteligentes descritos na literatura. Um desses classificadores é o Classificador Inteligente de Supernovas do tipo Ia (CIntIa) proposto por ([M6DOLO, 2016](#)), que usa Redes Neurais Artificiais (RNAs) para classificar os tipos principais: Ia, Ib, Ic e II. Outro classificador importante para o presente trabalho é o Sistema Fuzzy Avaliador de Supernovas (SUZAN), desenvolvido por ([ARANTES FILHO, 2018](#)). Ambos os classificadores foram desenvolvidos no âmbito da Pós-Graduação em Computação Aplicada do INPE.

CIntIa e SUZAN são parte de uma mesma iniciativa para fornecer sistemas de clas-

sificação automática de SNs para o projeto Kulun Dark Universe Survey Telescope (KDUST). O projeto está construindo um grande telescópio para o observatório de Kulun na Antártida, previsto para ser instalado em 2020 (ZHU et al., 2014). Devido a localização da estação e a grande quantidade de observações previstas para o KDUST, a automatização do processo de classificação se mostra essencial. No entanto, a CIntIa e a SUZAN não estão restritas ao projeto KDUST, é possível usá-las para classificar espectros obtidos por outros telescópios e outras aplicações.

O objetivo deste trabalho é propor o aperfeiçoamento do sistema CIntIa a fim de que ele aceite uma diversidade maior de dados e sua capacidade de generalizar seja expandida. Isso envolve, primeiramente, aceitar espectros com variação de comprimento de onda na faixa do visível (entre 4000 e 7000 Å), que é onde estão as linhas espectrais observadas para efetuar a classificação. O segundo ponto é o aceite de espectros capturados antes ou depois do brilho máximo (a classificação na fase de brilho máximo é o procedimento padrão para todos os classificadores em atividade), que é um avanço no emprego do classificador inteligente na prática, em campo de observação, conjuntamente com telescópios e espectroscópios. O uso de mais bases de dados, provenientes de equipamentos diferentes, também é uma contribuição fundamental para o cumprimento dos objetivos. Outra melhoria proposta para a CIntIa é o uso dos espectros duplamente filtrados e pré-processados de forma semelhante à SUZAN.

Aumentou-se em mais de 1300% (de 649 para 9156) a quantidade de espectros utilizados nas etapas de aperfeiçoamento da CIntIa, treinamento e testes, resultando no sistema que foi denominado CINTIA 2. Assim, a análise dos resultados apresenta-se mais robusta em relação à primeira versão da CIntIa. A nova análise ajudou a desenhar uma arquitetura de redes neurais binárias que faz com que o sistema não apresente classificações ambíguas, uma dificuldade presente na CIntIa. Um desdobramento natural da arquitetura foi a implementação de uma ferramenta desenvolvida nas linguagens Python e C++, permitindo assim a sua utilização por qualquer pessoa interessada.

Visando colaborar com os estudos de classificação de supernovas do tipo Ia em fases espectrais utilizando inteligência artificial, realizou-se experimentos, ainda iniciais, de classificação aplicando os métodos de redes neurais e K-Means. No entanto, ambos os métodos não apresentaram bons resultados.

Este trabalho está organizado da maneira como segue. O capítulo 2 trata brevemente da formação de SNs e conceitos importantes para o entendimento do trabalho e

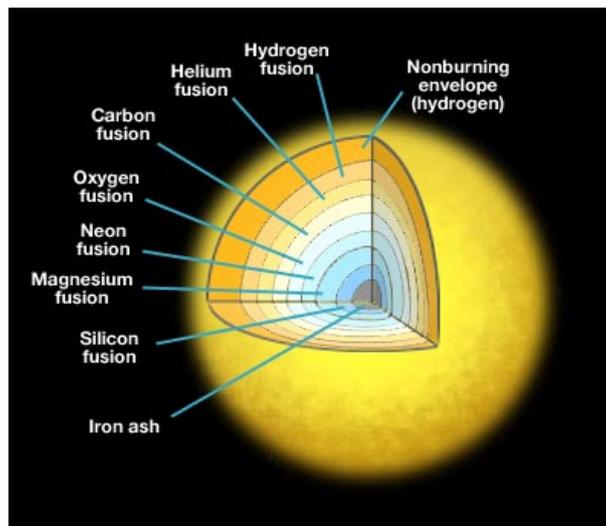
dos fundamentos da classificação de SNs. O capítulo 3 traz uma explanação sobre RNAs com ênfase no Perceptron de Múltiplas Camadas (MLP, do inglês Multi-Layer Perceptron) e na estratégia Um Contra Todos, usada para fundamentar a binarização das redes neurais. No capítulo 4, apresenta-se alguns classificadores automáticos descritos na literatura. O capítulo 5 aborda os dados usados, sua origem e pré-processamento. No capítulo 6, são descritas as métricas usadas para fazer a avaliação objetiva do classificador. Em seguida, no capítulo 7, detalha-se o desenvolvimento da CINTIA 2 até a sua arquitetura e no capítulo 8 o desenvolvimento do software CINTIA 2 é especificado. O capítulo 9 é a avaliação dos resultados obtidos. Em seguida, o capítulo 10 contém considerações sobre a classificação dos espectros de supernovas Ia por fase espectral. E finalmente, a conclusão faz uma síntese do que foi discutido no decorrer do trabalho.

2 SUPERNOVAS

Supernovas (SNs) são eventos caracterizados por grandes explosões que correspondem à fase final da vida de algumas estrelas. “[...] as grandes estrelas, ao sucumbir, superam galáxias inteiras em brilho. Seus clarões podem ser vistos por toda a extensão do Universo por alguns dias.” (DAMINELI; STEINER, 2010). Por ocorrerem apenas em grandes estrelas, ou seja, as muito massivas, SNs são eventos raros, aproximadamente uma ocorrência por galáxia a cada século, segundo (CARROLL; OSTLIE, 2007). Cada estrela segue uma sequência evolutiva que depende da massa que ela possui ao ser formada e se está sozinha ou em um sistema binário ou múltiplo. As estrelas com massa maior do que 10 massas solares (\odot) são as que geram eventos catastróficos. Esse tipo de explosão é conhecido como colapso de núcleo.

A explosão por colapso de núcleo acontece porque a estrela na sequência principal converte todo o H do núcleo em Hélio (He), transformando-se em gigante. Em seguida, consome Carbono (C) e Oxigênio (O) passando assim à fase de supergigante, quando consome Neônio (Ne), Magnésio (Mg) e Silício (Si) até restar apenas o núcleo de Ferro (Fe), as camadas da estrela são mostradas na Figura 2.1. Na última fase, a supergigante ejeta a maior parte da sua massa originando uma SN.

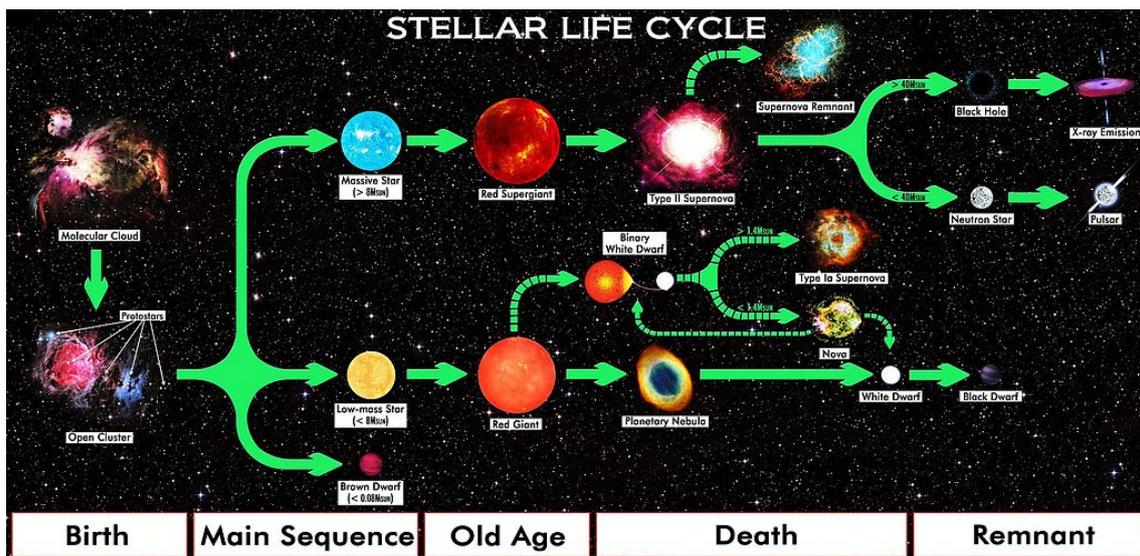
Figura 2.1 - Camadas de uma estrela massiva.



Fonte: Departamento de Física da Southern Methodist University (2018).

Além das SNs originadas pelo colapso do núcleo, existem as originadas por explosões termonucleares, que são aquelas criadas em sistemas binários, onde pelo menos uma das estrelas é uma anã branca com massa inicial superior a $0,8\odot$. A explosão é provocada pela instabilidade no núcleo devido ao acréscimo de massa na anã branca, que excede o limite de Chandrasekhar ($1,38\odot$). Segundo (WOOSLEY; WEAVER, 1986), a ignição de C ou He sob condições extremamente degeneradas queima uma massa substancial no núcleo que provoca a desintegração da estrela em altas velocidades. De acordo com (OLIVEIRA FILHO; SARAIVA, 2014), cerca de $0,6\odot$ é ejetada ao meio interestelar na forma de Fe, produzido durante a explosão, sendo esta a maior fonte de Fe conhecida. Na Figura 2.2 somos apresentados a um esquema que mostra a evolução estelar até a explosão da SN pelos dois processos discutidos acima e o que se forma de seu material remanescente.

Figura 2.2 - Fluxos de evolução da estrela de acordo com a sua massa.



Acima a formação de uma SN de colapso do núcleo e embaixo a formação de uma SN termonuclear. Também são representados os seus remanescentes.

Fonte: [Wikimedia \(2017\)](#).

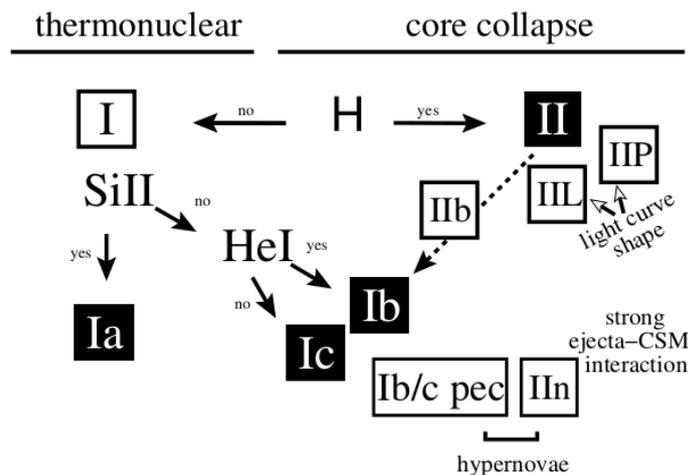
2.1 Classificação das Supernovas

“As supernovas são classificadas em dois tipos principais [...]: as SNs tipo I, que não apresentam H no espectro, e as SNs tipo II, que apresentam linhas de emissão

ou absorção de H no espectro [...]” (OLIVEIRA FILHO; SARAIVA, 2014). (CARROLL; OSTLIE, 2007) afirma que o tipo I pode ser subdividido de acordo com seu espectro. As supernovas do tipo Ia contam com uma forte presença de linhas de Si. As do tipo Ib apresentam linhas de He, enquanto as do tipo Ic não apresentam Si nem He.

As SNs do tipo Ia, de acordo com (TURATTO, 2003), são as originadas das explosões termonucleares. Por outro lado, as SNs de tipo Ib, Ic e II são provenientes do colapso do núcleo das estrelas massivas. A Figura 2.3 apresenta o esquema de classificação proposto por (TURATTO, 2003), onde é possível observar a divisão entre as SNs termonucleares e as de colapso de núcleo.

Figura 2.3 - Esquema de Turatto para classificação de supernovas.



Fonte: Turatto (2003).

Os tipos destacados em preto na Figura 2.3 (Ia, Ib, Ic e II) são os tipos que podem ser identificados a partir da análise do espectro de luz usando a técnica de espectroscopia. A espectroscopia é o estudo da luz, pela sua decomposição em comprimento de ondas ou frequências (cores). As três leis empíricas da espectroscopia foram formuladas por Gustav Kirchhoff (1824-1887) e são postuladas como segue, segundo (OLIVEIRA FILHO; SARAIVA, 2014):

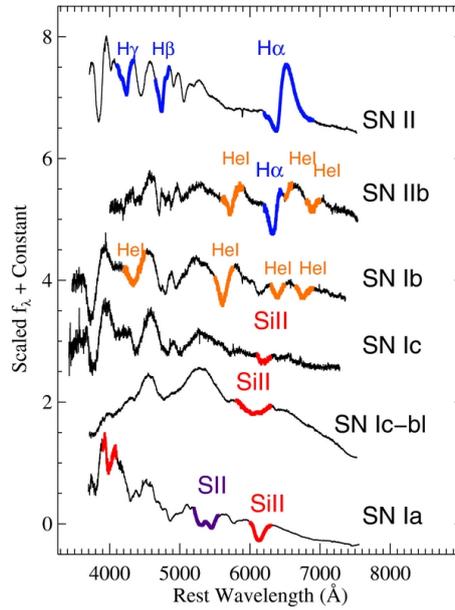
- Espectro contínuo: um corpo opaco quente, sólido, líquido ou gasoso, emite um espectro contínuo;

- Espectro de emissão: um gás transparente produz um espectro de linhas brilhantes (de emissão). O número e a cor (posição) dessas linhas depende dos elementos químicos presentes no gás; e
- Espectro de absorção: se um espectro contínuo passar por um gás a temperatura mais baixa, o gás frio causa a presença de linhas escuras (absorção). O número e a posição dessas linhas depende dos elementos químicos presentes no gás.

A classificação de SNs por espectroscopia, baseada na presença e ausência de elementos químicos, é normalmente rápida e pode ser feita assim que a SN é descoberta pelo telescópio, como esclarece (TURATTO et al., 2007). A Figura 2.4 mostra os aspectos mais comuns dos tipos principais de SNs que podem ser classificadas por espectroscopia. As regiões destacadas em cores por (MODJAZ et al., 2014) são aquelas em que há emissão ou absorção dos elementos químicos característicos formando picos e vales. Essas regiões são as que recebem maior atenção na classificação por RNAs proposta neste trabalho. “Os espectros das SNs Ia costumam exibir picos e vales largos atribuídos aos elementos O, Mg, Si, Enxofre (S), Cálcio (Ca), Fe e Cobalto (Co) neutros ou uma vez ionizados” (FILIPPENKO, 1997). Ainda segundo (FILIPPENKO, 1997), os elementos Si, S, e Ca estão presentes principalmente durante o brilho máximo da SN e o Fe após essa fase, aproximadamente duas semanas depois. A Figura 2.5 mostra espectros das SNs 1994D e 1987L, ambas do tipo Ia, capturados em dias diferentes. A forma dos espectros apresenta regiões semelhantes de picos e vales e sofrem alterações ao longo do tempo.

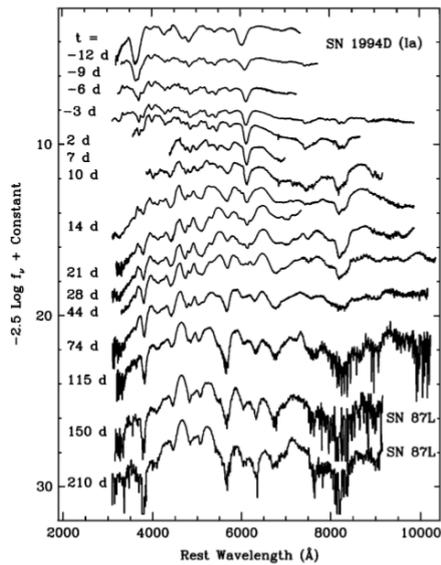
Cada valor de t expresso na Figura 2.5 é chamado de fase espectral e se refere ao dia (em relação ao dia em que a SN alcançou seu brilho máximo) em que aquele espectro foi auferido. O dia em que o brilho foi o mais intenso é a fase espectral 0 dias, se o espectro foi medido em dias anteriores, a fase tem valor negativo e em dias posteriores ela tem valor positivo. A fase espectral é importante para a classificação porque a forma do espectro muda com o tempo e apenas para alguns períodos existe consenso quanto à forma que o espectro de cada tipo deve ter. Por isso, a classificação automática tende a ser confiável apenas em fases mais próximas do 0 (brilho máximo) quando o padrão, principalmente para o tipo Ia, já é bem conhecido pelos especialistas. O inconveniente de classificar os espectros apenas na fase de brilho máximo é que pode ser necessário traçar a curva de luz da SN para inferir a fase de cada espectro, o que pode levar alguns meses, impedindo assim a classificação rápida em uma plataforma semelhante ao KDUST.

Figura 2.4 - Padrão dos espectros de SNs dos tipos Ia, Ib, Ic, Ic-bl, II e IIb.



Destaque em cores nos picos e vales que caracterizam cada um dos dos espectros.
Fonte: Modjaz et al. (2014).

Figura 2.5 - Evolução dos espectros das SNs 1994D e 1987L.



Fonte: Filippenko (1997).

3 REDES NEURAIIS ARTIFICIAIS

Neste capítulo, aborda-se as Redes Neurais Artificiais (RNAs) em detalhes. Na primeira seção traçamos um breve histórico. Na segunda seção são abordados os conceitos técnicos mais importantes para a compreensão deste trabalho. Em seguida, explana-se as características do Perceptron de Múltiplas Camadas (MLP, do inglês *Multi-Layer Perceptron*), que é o modelo usado majoritariamente neste trabalho, como será abordado em capítulos posteriores. E por último, apresenta-se o conceito de “*One Against All*” (Um Contra Todos) que justifica o uso de hierarquias de RNAs binárias.

3.1 Histórico

Uma RNA é uma rede que conecta unidades de processamento simples, os neurônios, e tem a capacidade de aprendizado por meio de exemplos. De acordo com (HAYKIN, 2001), uma RNA é semelhante ao cérebro em dois aspectos:

- O conhecimento é adquirido pela rede a partir do seu ambiente através de um processo de aprendizagem; e
- Forças de conexão entre neurônios conhecidos como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido.

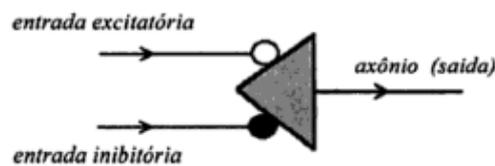
O modelo para RNA deriva do modelo biológico de funcionamento do cérebro que é composto por bilhões de neurônios. “Um neurônio é uma célula formada por três seções com funções específicas e complementares: corpo, dendritos e axônio. Os dendritos captam os estímulos recebidos em um determinado período de tempo e os transmitem ao corpo do neurônio, onde são processados.” (FERNEDA, 2006). Os neurônios transmitem estímulos uns aos outros através das sinapses e assim a informação é processada.

O primeiro modelo formal de um neurônio artificial foi desenvolvido por Warren McCulloch (1898 - 1969) e Walter Pitts (1923 - 1969) em 1943, como conta (DOMINGOS, 2017). As entradas do neurônio de McCulloch-Pitts (Figura 3.1) têm peso arbitrário e podem ser excitatórias (valor positivo) ou inibitórias (valor negativo). A saída binária é calculada pela soma ponderada das entradas com seus respectivos pesos, se o resultado é maior ou igual a um determinado limiar a saída é 1 (um), caso contrário a saída é 0 (zero), como esclarece (KOVÁCS, 2006).

Uma rede de neurônios de McCulloch-Pitts pode executar todas as operações comuns

de um computador, mas não é capaz de aprender. Então em 1958 o perceptron foi criado, os pesos variáveis entre as conexões dos neurônios possibilita o aprendizado neste modelo, como explica (DOMINGOS, 2017). O perceptron foi criado por Frank Rosenblatt (1928 - 1971), o nome deriva do interesse pela aplicação de seus modelos em tarefas perceptivas, como informa (DOMINGOS, 2017). O perceptron era um modelo simples e conseguiu se sair bem em tarefas como reconhecer caracteres e sons de voz. Para (KOVÁCS, 2006), Rosenblatt conseguiu dar dois saltos de qualidade com a criação do perceptron, primeiro o aumento em sua sensibilidade a estímulos, as saídas podem ser contínuas e não apenas binárias. A outra grande melhoria foi a introdução de uma lei de treinamento para o perceptron.

Figura 3.1 - Neurônio de McCulloch-Pitts.



Fonte: Adaptado de Kovács (2006).

Em 1969 foi publicado o livro “Perceptrons” de Marvin Minsky (1927-2016) e Seymour Papert (1928-2016) no qual eram apontadas as falhas do algoritmo perceptron. Então, até a década de 1980, as pesquisas em RNA foram drasticamente reduzidas e quase nada foi produzido no período até o advento do MLP. Nas palavras de (DOMINGOS, 2017), Minsky e Papert já admitiam, em seu livro, que camadas de neurônios interconectados poderiam aprender mais, no entanto não enxergavam uma maneira de simulá-la.

3.2 Conceitos de Redes Neurais Artificiais

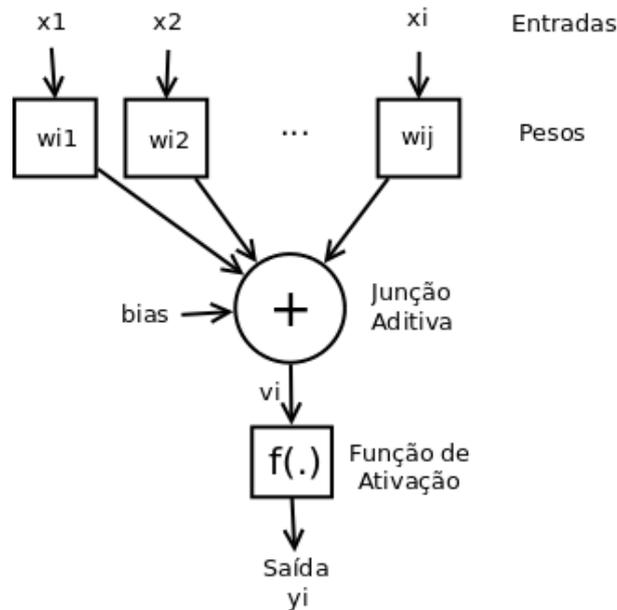
A estrutura básica formadora de uma RNA é o neurônio, a Figura 3.2 é a representação de seu modelo matemático. Os elementos que compõem o neurônio são os seguintes, de acordo com (HAYKIN, 2001):

- Peso sináptico: é um valor negativo ou positivo que representa a força de

conexão entre um dado de entrada e o neurônio. O primeiro índice de um peso se refere ao neurônio e o segundo índice é o mesmo índice da entrada;

- Somador: também chamado de junção aditiva, é responsável por somar todas as entradas ponderadas por seus pesos sinápticos;
- *Bias*: é um fator que aumenta ou diminui o valor da saída do somador;
- Função de ativação: é a função que define a saída do neurônio.

Figura 3.2 - Modelo de um neurônio artificial.



Fonte: Adaptado de Haykin (2001).

De acordo com a argumentação de (HAYKIN, 2001), as Equações 3.1 e 3.2 descrevem o comportamento matemático do neurônio artificial representado na Figura 3.2.

$$v_i = \sum_{k=1}^m w_{ik}x_k \quad (3.1)$$

$$y_i = \phi(v_i + b_i) \quad (3.2)$$

No que concerne à função de ativação, as funções mais usadas são a função limiar, a função linear, a função sigmóide e a função tangente hiperbólica (Tabela 3.1). A função de ativação deve ser escolhida de acordo com a natureza do problema de aprendizagem tratado.

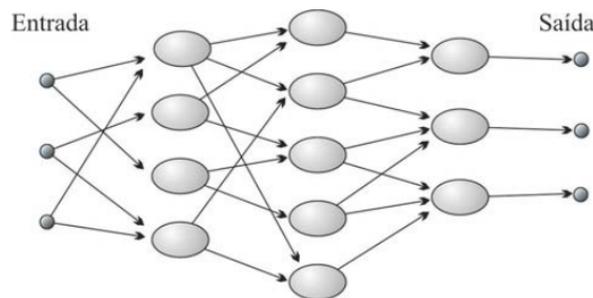
Tabela 3.1 - Funções de ativação mais usadas.

Tipo de Função	Função
Limiar	$\varphi(v) = 1, v \geq 0$ ou $\varphi(v) = 0, v < 0$
Linear	$\varphi(v) = 1, v \geq 1/2$ ou $\varphi(v) = v, -1/2 < v < 1/2$ ou $\varphi(v) = 0, v \leq -1/2$
Sigmóide	$\varphi(v) = 1/(1 + \exp(-av))$
Tangente Hiperbólica	$\varphi(v) = \tanh(v)$

Fonte: Adaptado de Quiles (2004).

A conexão entre mais de um neurônio molda a arquitetura de uma RNA. A ligação entre os neurônios pode ser feita em apenas uma camada, na qual as entradas são multiplicadas pelos pesos, passam pelo somador e logo em seguida pela função de ativação. Ou a RNA pode ter mais de uma camada (Figura 3.3), assim a saída de uma camada é usada como entrada para os neurônios da camada posterior sendo propagada até a saída.

Figura 3.3 - Representação de uma RNA com mais de uma camada.



Fonte: Farneda (2006).

O aprendizado da RNA, por sua vez, pode ser do tipo supervisionado ou não-supervisionado. No aprendizado supervisionado, em oposição ao não-supervisionado,

um agente externo apresenta à RNA alguns padrões de entrada e seus correspondentes padrões de saída. Portanto, é necessário ter um conhecimento prévio do comportamento que se deseja ou se espera da rede, como esclarece (FERNEDA, 2006). O aprendizado depende de um algoritmo que modifica os pesos de acordo com as saídas da RNA em cada sessão de treinamento. O algoritmo de atualização dos pesos, conhecido como retropropagação (*backpropagation*), é tratado na próxima seção onde o MLP é abordado com mais detalhes.

3.3 Perceptron de Múltiplas Camadas (MLP)

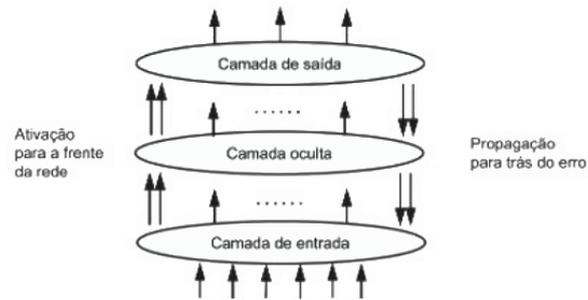
Uma rede MLP consiste de um conjunto de unidades que constituem a camada de entrada, uma ou mais camadas ocultas e uma camada de saída, como descreve (HAYKIN, 2001). A entrada se propaga camada por camada até chegar a saída. Ainda para (HAYKIN, 2001), uma rede do tipo MLP apresenta três características básicas:

- O modelo de cada neurônio da rede possui uma função de ativação não-linear, uma das funções mais utilizadas é a função sigmóide;
- A rede tem uma ou mais camadas ocultas, ou seja, camadas intermediárias entre a entrada e a saída da rede; e
- A rede exhibe um alto grau de conectividade entre os neurônios.

O aprendizado no MLP é comumente realizado com o algoritmo de retropropagação que é baseado na regra de aprendizagem por correção de erro. O aprendizado por correção de erro funciona basicamente adquirindo o sinal de erro da saída e a partir dele os pesos são corrigidos até chegar em um valor aceitável. O algoritmo de retropropagação foi criado em 1986 por David Rumelhart (1942-2011) e foi o responsável pela ressurgimento do interesse nas RNAs. “A abordagem adotada pelo algoritmo [retropropagação] consiste em iniciar na camada de saída e propagar o erro retroativamente através das camadas ocultas.” (LUGER, 2013). A Figura 3.4 é uma representação da retropropagação em um MLP com uma camada oculta.

O treinamento com retropropagação usa a regra delta generalizada. O cálculo do fator delta de cada neurônio depende se o neurônio está na camada de saída ou em uma camada intermediária. As Equações 3.3 e 3.4 são as fórmulas para calcular o erro e o delta, respectivamente, em neurônios da camada de saída. Enquanto a Equação 3.5 é a fórmula para calcular o delta em neurônios de camadas ocultas. As equações são as mesmas apresentadas por (HAYKIN, 2001).

Figura 3.4 - Representação da retropropagação em um MLP com uma camada oculta.



Fonte: Adaptado de Luger (2013).

$$e_i = d_i - y_i \quad (3.3)$$

Onde:

- d_i : resultado esperado do neurônio i ;
- y_i : resultado verificado na saída do neurônio i .

$$\delta_i = e_i \dot{\varphi}_i(v_i) \quad (3.4)$$

Onde:

- e_i : erro da saída do neurônio i ;
- $\dot{\varphi}_i(v_i)$: derivada da função de ativação.

$$\delta_i = \dot{\varphi}_i(v_i) \sum_k \delta_k w_{ik} \quad (3.5)$$

Onde:

- $\dot{\varphi}_i(v_i)$: derivada da função de ativação;

- $\sum_i \delta_i w_{ik}$: somatório ponderado dos deltas calculados para os neurônios da camada posterior à qual o neurônio i está conectado.

A Figura 3.5 explicita os valores que estão envolvidos no cálculo da correção dos pesos no algoritmo de retropropagação. Resumindo, o algoritmo segue dois passos principais. O primeiro passo é a propagação para a frente, durante a qual os pesos não são alterados e o sinal de erro é calculado ao fim da propagação. No segundo passo, os deltas derivados do sinal de erro são propagados para trás e os pesos são atualizados. Cada vez que esses dois passos são executados se dá uma sessão de treinamento, também chamada de época de treinamento. São executadas quantas épocas forem necessárias até que a rede tenha aprendido suficientemente de acordo com algum parâmetro chamado condição de parada.

Figura 3.5 - Como calcular a correção dos pesos no algoritmo de retropropagação.

$$\begin{pmatrix} \text{Correção} \\ \text{de peso} \\ \Delta w_{ji}(n) \end{pmatrix} = \begin{pmatrix} \text{Parâmetro da} \\ \text{taxa de aprendizagem} \\ \eta \end{pmatrix} \cdot \begin{pmatrix} \text{Gradiente} \\ \text{local} \\ \delta_j(n) \end{pmatrix} \cdot \begin{pmatrix} \text{sinal de entrada} \\ \text{do neurônio } j \\ y_i(n) \end{pmatrix}$$

Fonte: Haykin (2001).

3.4 Estratégia Um Contra Todos para Problemas Multi-classe

De acordo com (OONG; ISA, 2012), em reconhecimento de padrões, usar um grupo de redes neurais para resolver um problema é mais efetivo do que usar uma rede neural única. A estratégia conhecida como Um Contra Todos é usada para solucionar problemas chamados de problemas multi-classe. Um problema multi-classe é aquele em que um objeto pode ser classificado apenas como uma classe, no entanto há mais de duas opções de classe. Duas fases são executadas para que se obtenha uma classificação usando essa estratégia: decomposição e reconstrução.

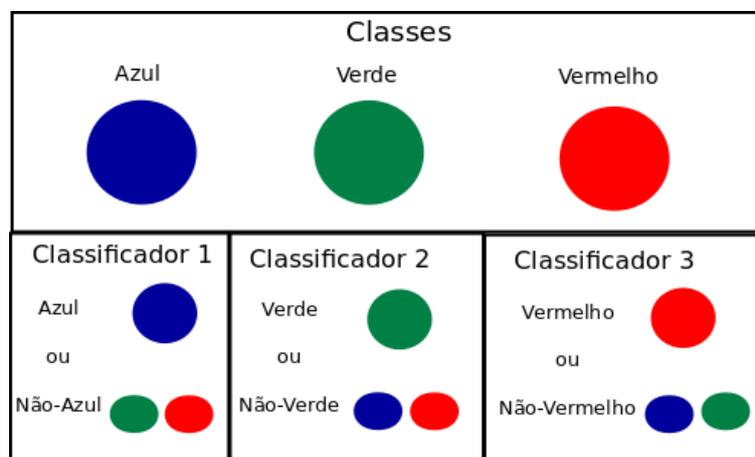
3.4.1 Fase de Decomposição

No domínio dos MLPs, muitos problemas multi-classe se tornam complexos e a separação das classes não é possível, logo a aprendizagem da rede é comprometida. Portanto, muitos pesquisadores optam por decompor o problema em subproblemas

mais simples. Assim, segundo (MENCIA; FURNKRANZ, 2008), um conjunto de treino com K possíveis classes é decomposto em K conjuntos binários de teste que são usados para treinar K classificadores binários. Como afirma (PIMENTA, 2004), cada conjunto de treino binário será composto pelos exemplos da classe principal e pelos exemplos de todas as outras classes, neste caso, as diversas classes são substituídas por uma única que é a negação da principal. Assim, para cada rede neural binária uma classe é considerada positiva e deve apresentar a saída 1, enquanto todas as outras classes são negativas e devem ter como saída 0 (as saídas 1 e 0 foram adotadas neste trabalho, mas são descritas com outros valores em outras fontes).

A Figura 3.6 apresenta um esquema que exemplifica a decomposição de um problema multi-classe, que deseja classificar a cor de cada bola como azul, verde ou vermelha. O problema com 3 classes é decomposto para 3 classificadores que apresentem resultados binários, onde cada uma das classes é considerada classe principal uma vez enquanto as outras formam uma classe geral antagonista.

Figura 3.6 - Fase de decomposição de um problema multi-classe que deve classificar bolas nas cores azul, verde ou vermelha.



Fonte: Produção da autora.

3.4.2 Fase de Reconstrução

Após a decomposição do problema deve ser considerada a decisão final sobre a qual classe pertence o objeto a ser classificado, esta etapa é chamada de fase de reconstrução. Na fase de reconstrução, agregam-se todas as predições feitas pelos classifi-

cadres binários para obter a predição final, como esclarece (PIMENTA, 2004). Ou seja, conseqüentemente à classificação feita por cada RNA binária, uma estratégia deve ser adotada para escolher a solução final mais apropriada.

Estratégias como a votação direta distribuída são empregadas para fazer a reconstrução. No algoritmo mais tradicional da votação distribuída, cada objeto é classificado por todos os classificadores, se a saída for positiva para uma classe ela recebe um voto, caso o classificador dê uma resposta negativa todas as outras classes recebem um voto. Ao final, os votos são contados e a classe com a maior quantidade de votos ganha, como esclarece (PIMENTA, 2004). Empregar esse método é simples, porém pode haver casos de empate e uma estratégia auxiliar deve ser adicionada.

(MENCIA; FURNKRAZ, 2008) mencionam a estratégia de ranking de classes, que pode ser aleatório ou deliberadamente escolhido. No ranking de classes, cada classe recebe um grau de precedência, a classe de maior precedência que for a saída de algum dos classificadores é considerada como a classe final do objeto. Neste trabalho, é adotado o ranking de classes e a atribuição de precedências foi feita de acordo com os testes empíricos dos treinamentos efetuados em cada rede neural binária. Ou seja, é praticado um ranking de classes no qual as precedências são escolhidas baseadas nos resultados dos testes efetuados em cada RNA após os treinamentos. A precedência é implementada por meio de uma hierarquia, o classificador (a partir do segundo) é acionado apenas se o seu antecessor não retornou uma resposta positiva. As fases de decomposição e reconstrução aplicadas neste trabalho serão melhor descritas no Capítulo 7.

4 CLASSIFICADORES AUTOMÁTICOS DE SUPERNOVAS

A classificação automática de SNs pode ser realizada de dois modos, analisando o espectro de luz (espectroscopia) ou a curva de luz (fotometria). A classificação por espectroscopia costuma ser imediata porque a análise pode ser feita com apenas um espectro. Com dados de fotometria, a classificação só pode ser feita muitos dias depois da explosão da SN porque é necessário estabelecer o perfil temporal da intensidade do brilho. Portanto, o foco deste trabalho é a classificação usando dados espectroscópicos. Foram selecionados seis classificadores de SNs que usam espectro de luz para abordagem neste capítulo.

4.1 Supernova Identification (SNID)

O SNID é uma ferramenta desenvolvida por (BLONDIN; TONRY, 2007) com a finalidade de determinar idade, *redshift* e classificar SNs. A identificação das SNs, que resulta em sua classificação é feita usando técnicas estatísticas. O espectro que se deseja classificar, cujo *redshift* deve ser conhecido, é correlacionado com outros espectros previamente identificados, o espectro recebe a mesma classificação daquele com quem mais se correlacionou.

Foram usados 879 espectros de 65 SNIa, 322 de 19 SNIb/c e 353 de 10 SNII no desenvolvimento do classificador. Os espectros foram previamente pré-processados para serem submetidos à correlação com fluxo normalizado entre 0 e 1 e filtrados. Os resultados da classificação do SNID discutidos pelos autores são pautados em cinco casos, que são:

- Distinção entre SNIa similares a 1991T e outras SNIa;
- Distinção entre SNIb/c e SNIa em altos *redshifts*;
- Identificação de SNIa peculiares;
- Distinção entre SNIb e SNIc;
- Distinção entre SNIb, SNII e SNIb.

Os autores consideram que o classificador é bem sucedido, principalmente nos três primeiros casos, que envolvem SNIa e esse comportamento é creditado em parte a maior quantidade de espectros desse tipo no banco de espectros. No entanto, a ferramenta carece de melhorias no reconhecimento de outros tipos, principalmente

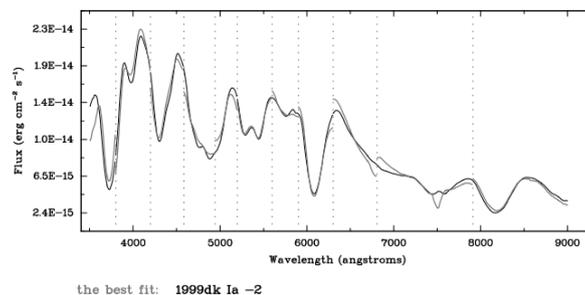
para as chamadas SNs peculiares. O Harvard-Smithsonian Center for Astrophysics (CfA) mantém um acervo online de espectros de SNs classificados pelo SNID até o ano de 2015.

4.2 Generic Classification Tool (GELATO)

O GELATO, apresentado na tese de doutorado (HARUTYUNYAN, 2008), é um classificador automático de SNs que pretende mitigar a subjetividade das classificações feitas por especialistas humanos. A ferramenta usa um método matemático para comparar novos espectros com outros previamente classificados de um banco de dados.

Primeiramente os espectros são pré-processados com correção do *redshift*, suavização, reamostragem e a divisão em 11 intervalos do comprimento de onda. A comparação entre os espectros é feita em cada um desses intervalos. A ferramenta computa para cada intervalo a distância relativa entre o novo espectro e todos os espectros do banco de dados. Por fim, calcula-se a média das distâncias relativas, o menor valor se refere ao espectro mais parecido, portanto os espectros recebem a mesma classificação. A Figura 4.1 mostra a classificação da SN2002bo cuja semelhança maior é com o espectro da SN1999dk que é do tipo Ia. As linhas pontilhadas são os intervalos em que o espectro é dividido. O autor relata que o GELATO é usado no dia-a-dia do seu grupo de pesquisa e que dá bons resultados, mas não são apresentadas métricas de avaliação no trabalho para comparações futuras. O GELATO em sua versão web está disponível para uso gratuito em <https://gelato.tng.iac.es/>

Figura 4.1 - Resultado da classificação da SN 2002bo pelo GELATO.



Fonte: Harutyunyan (2008).

4.3 Classificador Inteligente de supernovas do tipo Ia (CIntIa)

A CIntIa é uma classificadora inteligente de SNs dos tipos Ia, Ib, Ic e II e foi desenvolvida por (M6DOLO, 2016) no âmbito da pós-graduação da Computação Aplicada do INPE. A ferramenta usa 4 RNAs para reconhecer cada um dos tipos principais de SNs e apresenta ótimos resultados na classificação do tipo Ia.

A CIntia usa 559 espectros de 192 SNIa, 33 espectros de 12 SNIb, 44 espectros de 12 SNIc e 13 espectros de 5 SNII totalizando 649 espectros. Todos os espectros selecionados estão entre 3 dias antes (-3) e 7 dias depois (+7) do brilho máximo das SNs. Os espectros passam por um pré-processamento antes de serem submetidos à RNA, que inclui: correção do *redshift*, suavização, interpolação, e normalização do fluxo.

A etapa seguinte é extrair intervalos do espectro que são usados como entradas nas RNAs. Cada intervalo corresponde à região onde se encontram ou não os elementos H, Si, S e He. As regiões escolhidas são as mais usadas pelos especialistas humanos para efetuar a classificação, são elas:

- 5000 a 6500 Å: para classificar em tipo Ia ou não;
- 5500 a 7000 Å: para classificar em tipo Ib ou não;
- 5500 a 6500 Å: para classificar em tipo Ic ou não;
- 4000 a 5000 Å e 6000 a 7000 Å: para classificar em tipo II ou não.

Para treinar a RNA foram selecionados 60% dos espectros. Foram testados dois conjuntos, cada qual com 20% dos espectros, cujos melhores resultados são apresentados na Tabela 4.1. A classificação do tipo Ia apresenta excelente resultado e com exceção do tipo Ib as outras RNAs apresentam uma concordância com a classificação real de moderada a boa, no entanto a quantidade de espectros usada no teste é pequena.

Tabela 4.1 - Avaliação das 4 RNAs que compõem o sistema CIntIa.

	Tipo Ia	Tipo Ib	Tipo Ic	Tipo II
Acurácia	0.99	0.97	0.97	0.98
Precisão	0.99	1	1	0.33
Recall	1	0.2	0.5	1
Kappa	0.95	0.32	0.65	0.49

Fonte: Adaptado de [Módolo \(2016\)](#).

4.4 Dimensionality Reduction and Clustering for Unsupervised Learning in Astronomy (DRACULA)

O DRACULA é um sistema desenvolvido por ([SASDELLI et al., 2016](#)) para identificar subtipos do tipo Ia. Para isso o DRACULA primeiro reduz a dimensionalidade dos dados e em seguida usa aprendizado não-supervisionado para realizar a classificação.

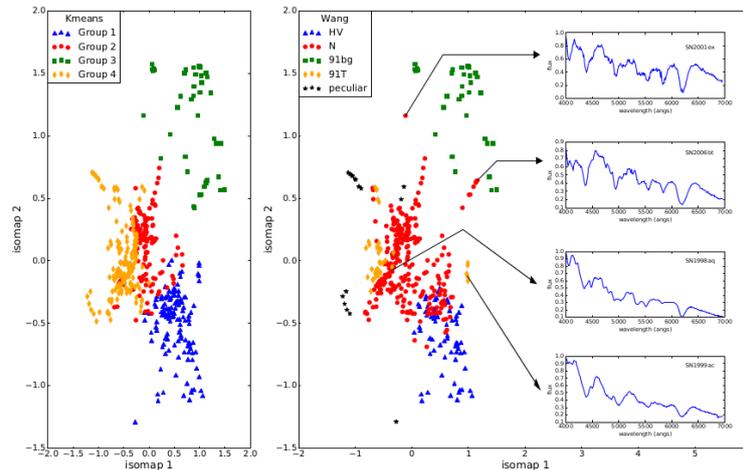
Foram usados 3677 espectros do tipo Ia na fase de redução de dimensionalidade que foi realizada com a técnica de *Deep Learning* (DL). Espectros dentro do intervalo 4000 a 7000 Å foram inicialmente interpolados gerando espectros com 300 pontos cada. Os 300 pontos foram transformados em apenas 4 características pelo algoritmo de DL usado. Em seguida, dos 3677 espectros foram selecionados 486 que foram auferidos entre -3 e +3 dias do brilho máximo das SNs.

Os 486 espectros selecionados após a redução da dimensionalidade foram submetidos ao algoritmo de aprendizado não-supervisionado K-Means. O K-Means divide as amostras em classes de acordo com semelhanças que elas possuam entre si sem que seja fornecido nenhum exemplo. O parâmetro necessário para a execução do algoritmo é o número de classes, os autores optaram por quatro classes de acordo com a divisão em subtipos do tipo Ia proposta por ([WANG et al., 2009](#)).

Os resultados apresentados mostram que usar DL, nesse caso, é mais efetivo do que usar outros algoritmos tradicionais de redução da dimensionalidade, como o *Principal Component Analysis*. A classificação das SNs não ficou totalmente em acordo com a proposta por ([WANG et al., 2009](#)) e não há métricas de avaliação no trabalho suficientes para a discussão. A Figura 4.2 mostra a divisão em classes feita pelo DRACULA em comparação com as classes de ([WANG et al., 2009](#)), foi usado o algoritmo Isomap que reduziu a dimensionalidade de quatro para duas facilitando assim a visualização. O DRACULA está disponível como um pacote para linguagem

de programação Python em <https://github.com/COINtoolbox/DRACULA>

Figura 4.2 - Resultado da classificação feita pelo DRACULA (à esquerda) e as subclasses do tipo Ia propostas por (WANG et al., 2009) (à direita).



Fonte: Sasdelli et al. (2016).

4.5 Quantitative Classification of Type I Supernovae (Quantitative)

A proposta de (SUN; GAL-YAM, 2017) é a classificação quantitativa dos subtipos do tipo I de SNs. Na classificação quantitativa, um tipo é identificado de acordo com características que ele possui e não pelas características que ele não possui. Por exemplo, nos métodos tradicionais o tipo Ic é classificado pela ausência de Si e He, enquanto na classificação quantitativa ele deve ser classificado pela presença de alguma característica definida. O método proposto por (SUN; GAL-YAM, 2017) para classificar quantitativamente é medir a profundidade de linhas de absorção dos elementos Si II (λ 6150 Å) e O I (λ 7774 Å).

Foram usados 146 espectros do tipo Ia, 12 do tipo Ib, 19 do tipo Ic e 4 do tipo Ib/c todos de SNs em *redshift* baixo (<0.1). Os espectros passaram por pré-processamento, que consistiu em: correção do *redshift* e suavização com o filtro de Savitzky-Golay. A segunda etapa no tratamento dos dados foi encontrar o pseudo-contínuo que delimita as regiões de absorção a ser analisadas, esta etapa foi realizada manualmente por apenas um dos autores. A região foi medida e em seguida a profundidade da região foi calculada. As profundidades medidas resultaram em um critério de classificação

desenvolvido pelo autores, que consiste em:

- SNIa: $a(\lambda 6150 \text{ \AA}) > 0.35$;
- SNIb: $a(\lambda 6150 \text{ \AA}) < 0.35$ e $\frac{a(\lambda 6150 \text{ \AA})}{a(\lambda 7774 \text{ \AA})} > 1$;
- SNIc: $a(\lambda 6150 \text{ \AA}) < 0.35$ e $\frac{a(\lambda 6150 \text{ \AA})}{a(\lambda 7774 \text{ \AA})} < 1$;

Onde:

- $a(\lambda 6150 \text{ \AA})$: profundidade da região de absorção do Si II;
- $a(\lambda 7774 \text{ \AA})$: profundidade da região de absorção do O I.

O resultado da profundidade das regiões foi comparado com o apresentado por (SILVERMAN et al., 2012a) e foi constatado que as medidas resultantes foram um pouco mais baixas do que a dos autores consultados, mas apenas 5% do total apresentaram grande discrepância. Assim 119 medidas do Si II e 40 medidas do O I foram concordantes. Apesar de não saberem explicar a implicância física de $\frac{a(\lambda 6150 \text{ \AA})}{a(\lambda 7774 \text{ \AA})}$, os autores afirmam que a maioria das SNs do tipo I podem ser classificadas pelos critérios apresentados por eles, exceto por: 2002cx Ia-pec, Ic-BL e as semelhantes à SNIa 1991T.

4.6 Sistema Fuzzy Avaliador de Supernovas (SUZAN)

A SUZAN, apresentada por (ARANTES FILHO, 2018) é um sistema classificador de SNs em seus tipos principais Ia, Ib, Ic e II. O sistema usa lógica nebulosa. Assim como a CIntIa, a SUZAN foi realizada no programa de pós-graduação de Computação Aplicada do INPE e foi desenvolvida para atuar em redundância com a CIntIa.

No total foram usados 3697 espectros de 588 SNs diferentes, dos quais 3082 espectros são de SNIa, 217 do tipo Ib, 282 do tipo Ic e 116 do tipo II. Os espectros passaram pela etapa de pré-processamento que consistiu na Dupla-Filtragem usando o filtro de Savitzky-Golay, ou seja, os espectros foram filtrados duas vezes seguidas após o *redshift* ter sido ajustado.

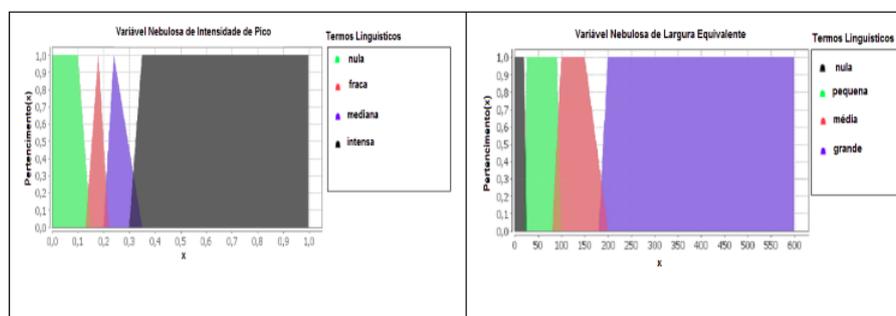
Segundo o autor, o modelo de classificação da SUZAN foi baseado no modelo de classificação espectral em que o astrônomo identifica no espectro as principais linhas de absorção e emissão e relaciona o elemento correspondente. Os elementos que são

buscados são H, He, Si e S. A SUZAN extrai características dos espectros para em seguida submetê-los ao crivo das regras nebulosas criadas com conhecimento especialista. As características são obtidas em três etapas:

- Busca de linhas de emissão e absorção;
- Cálculo da distância relativa das linhas encontradas em relação às linhas teóricas;
- Abstração dos parâmetros nas características Intensidade de Pico e Largura Equivalente.

A Figura 4.3 mostra as funções de pertencimento e as variáveis linguísticas associadas às variáveis de entrada Intensidade de Pico e Largura Equivalente. Essa etapa seleciona apenas as linhas de emissão e absorção (linhas candidatas) que se adequam aos padrões esperados para as linhas dos elementos químicos buscados. Então é realizada uma nova etapa de processamento por regras nebulosas que avaliam a distância relativa de cada linha candidata (Figura 4.4). Assim, ao final do processo é verificada a diversidade de elementos em cada espectro e aplicada a classificação proposta por (TURATTO, 2003).

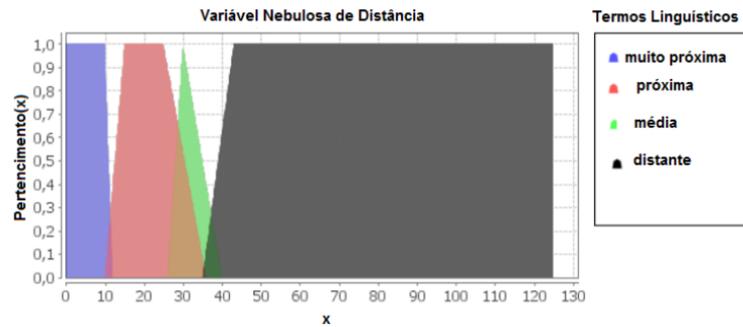
Figura 4.3 - Funções de pertencimento e variáveis linguísticas das variáveis de entrada Intensidade de Pico e Largura Equivalente.



Intensidade de Pico à esquerda e Largura Equivalente à direita.

Fonte: Adaptado de Arantes Filho (2018).

Figura 4.4 - Função de pertencimento e variáveis linguísticas da variável de entrada Distância Relativa.



Fonte: Arantes Filho (2018).

O melhor desempenho da SUZAN acontece na classificação de espectros de SNIa na fase de brilho máximo, assim os elementos como o S e o Si são identificados pelas regras nebulosas, é neste período que os astrônomos especialistas classificam os espectros de SNIa. Dessa forma, o autor propõe uma separação de espectros de SNIa apenas na fase de brilho máximo (entre -2.5 dias a +2.5 dias), período proposto por (BLONDIN et al., 2011) que considera ser o período em que as características do Si e do S estão mais evidentes.

5 DADOS

Os dados usados neste trabalho foram obtidos no The Open Supernova Catalog (TOSC), catálogo de SNs que reúne dados de 17 bases de espectros mais contribuições feitas por pesquisadores individualmente (GILLOCHON et al., 2017). O acesso ao catálogo é gratuito, a Tabela 5.1 lista as principais referências dos bancos de espectros catalogados pelo TOSC.

Tabela 5.1 - Bancos de espectros compilados no TOSC e suas principais referências.

Banco de espectros	Principais referências
Asiago SN Catalog	(BARBON et al., 2009)
CCCP	(GAL-YAM et al., 2004)
CPCS	(CAMBRIDGE PHOTOMETRY CALIBRATION SERVER, 2019)
CSP	(HAMUY et al., 2006)
CfA	(MATHESON et al., 2008); (BLONDIN et al., 2011); (BLONDIN et al., 2012); (MODJAZ et al., 2014)
Gaia Alerts	(CAMPBELL et al., 2014)
Latest Supernovae	(LATEST SUPERNOVAE, 2019)
SNF	(ALDERING et al., 2002)
OGLE-IV	(WYRZYKOWSKI et al., 2014)
Pan-STARRS	(CHAMBERS; WERNER, 2014); (MAGNIER et al., 2016); (FLEWELLING et al., 2016)
SDSS	(ABAZAJIAN et al., 2003)
Sternberg Catalogue	(STERNBERG ASTRONOMICAL INSTITUTE SUPERNOVA CATALOGUE, 2019)
SNHunt	(HOWERTON, 2017)
SNLS	(ASTIER et al., 2006)
SUSPECT	(RICHARDSON et al., 2002)
SNDB	(SILVERMAN et al., 2012b)
WISeREP	(YARON; GAL-YAM, 2012)

Fonte: Produção da autora.

A obtenção dos espectros, dados usados como entradas do classificador, foi feita seguindo alguns passos, pois além dos espectros também são necessárias outras informações para que a classificação seja efetuada. Enumera-se as etapas executadas para baixar o conjunto de dados (os espectros foram baixados no dia 30 de abril de 2018). As etapas a, b, c, d e f foram realizadas manualmente, enquanto as etapas e, g e h foram feitas por meio de *scripts* programados com a linguagem Python.

- a) Baixar um arquivo com os metadados das SNs presentes no catálogo, as informações dos metadados utilizadas neste trabalho são: nome da SN, nome da galáxia onde a SN está localizada, tipo da SN e quantidade de espectros;
- b) Excluir do arquivo de metadados de SNs todos os registros de SNs que não possuem espectros catalogados;
- c) Excluir do arquivo de metadados de SNs todos os registros de SNs que não sejam dos tipos Ia, Ib, Ic e II, inclusive os subtipos;
- d) Baixar um arquivo com os metadados de todas as galáxias presentes no catálogo e assim obter os valores de *redshift*;
- e) Adicionar ao arquivo de metadados de SNs os valores de *redshift*;
- f) Excluir todos os registros de SNs cujo valor de *redshift* não seja conhecido;
- g) Baixar os espectros de todas as SNs que continuaram no arquivo de metadados;
- h) Criar arquivo com dados de fase espectral de todos os espectros.

5.1 Limpeza do Conjunto de Dados

Um espectro é representado por um arquivo, com extensão .csv, que contém duas colunas, na primeira coluna estão os valores de comprimento de onda e na segunda, os valores de fluxo correspondentes à cada comprimento. Assim, a limpeza do conjunto de dados inicialmente baixado é necessária para eliminar arquivos inconsistentes, ou seja, aqueles que não são apropriados para submeter aos algoritmos de pré-processamento e classificação. Foram descartados todos os arquivos vazios, arquivos com todos os valores de fluxo iguais (formando apenas uma linha reta) e arquivos sem informação de fase espectral. A quantidade inicial de dados era 12902 espectros de 4883 SNs. Após a limpeza, a quantidade de SNs diminuiu para 4392 e os espectros foram reduzidos para 12271.

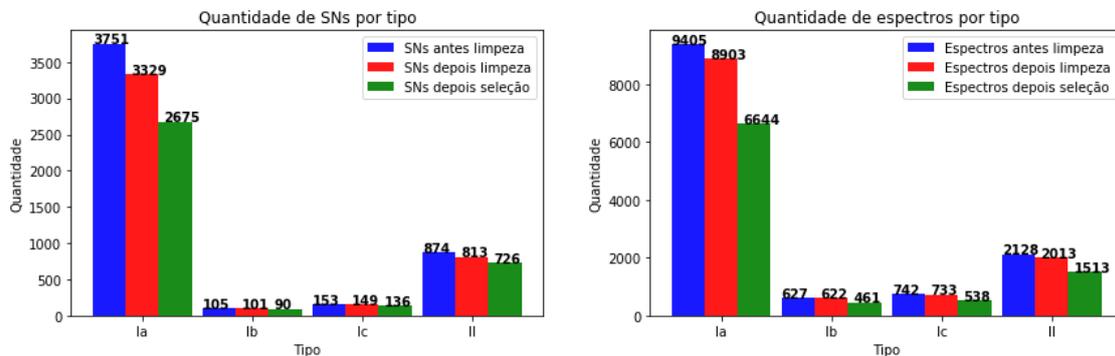
5.2 Seleção de Intervalo de Comprimento de Onda

A quantidade de espectros efetivamente utilizada nas etapas de treinamento e teste do classificador é menor que o total resultante da limpeza de dados. Após a etapa de limpeza dos dados, foi aplicado um critério de seleção de espectros aptos para a

classificação que consiste em excluir aqueles que não possuem intervalo de comprimento de onda entre 4000 e 7000 Å. A escolha desse intervalo decorre da observação das linhas espectrais importantes para a avaliação dos espectros, as mesmas usadas pelos especialistas, as quais estão dentro da faixa de luz visível.

A Figura 5.1 mostra o gráfico da quantidade de SNs e de espectros por tipo antes e depois da limpeza do conjunto e depois da aplicação desse critério. A quantidade de dados do tipo Ia é bastante superior, o que é benéfico para a melhoria da CIntIa, que prioriza a classificação das SNs Ia. A Figura 5.2 exibe os histogramas das fases espectrais por tipo de SN. A análise dos histogramas deixa claro que a maioria dos espectros, de todos os tipos, tem fase espectral em torno do brilho máximo, o que reforça o padrão buscado pelo algoritmo de aprendizagem que está sendo usado.

Figura 5.1 - Quantidade de SNs (esquerda) e de espectros (direita) por tipo antes da limpeza, depois da limpeza do conjunto de dados e após aplicar o critério de seleção de intervalo de comprimento de onda.



Fonte: Produção da autora.

5.3 Pré-Processamento

Após a limpeza do conjunto de dados, os espectros que permaneceram no conjunto passaram por uma etapa de pré-processamento que visa eliminar ruídos, fontes de interferências, que possam prejudicar a aprendizagem dos padrões realizada pelo classificador.

5.3.1 Ajuste do Redshift

O *redshift* é o deslocamento em direção a comprimentos de onda longos (vermelhos), como define a (BRITANNICA, 2018). O ajuste do *redshift* é recomendado para que o espectro seja analisado como se os gases da explosão correspondente estivessem em repouso. Dessa forma, as linhas espectrais do evento podem ser comparadas com as linhas medidas em laboratório. A correção é feita com a Equação 5.1.

$$\lambda_0 = \frac{\lambda}{z + 1} \quad (5.1)$$

Onde:

- λ_0 : comprimento de onda do objeto em repouso;
- λ : comprimento de onda observado;
- z : *redshift*.

5.3.2 Dupla-Filtragem com Filtro Savitzky-Golay

Para realizar a Dupla-Filtragem, inicialmente os espectros, os quais já passaram pelo ajuste de *redshift*, são reamostrados em 1000 pontos cada, por uma interpolação linear simples que começa no ponto inicial e termina no ponto final de cada espectro. Em seguida, cada espectro tem seu fluxo normalizado com valores entre 0 e 1. Então, os espectros são submetidos à Dupla-Filtragem, a fim de reduzir ruídos e remover inconsistências, como esclarece (ARANTES FILHO, 2018), que desenvolveu o método para ser usado na SUZAN.

A filtragem é realizada pelo filtro de Savitzky-Golay (SG), filtro desenvolvido especialmente para ser usado em dados espectroscópicos, dessa forma ele se encaixa bem no problema que tratamos neste trabalho. O filtro de SG se comporta essencialmente como um método de média ponderada e segue a Equação 5.2, segundo (SAVITZKY; GOLAY, 1964).

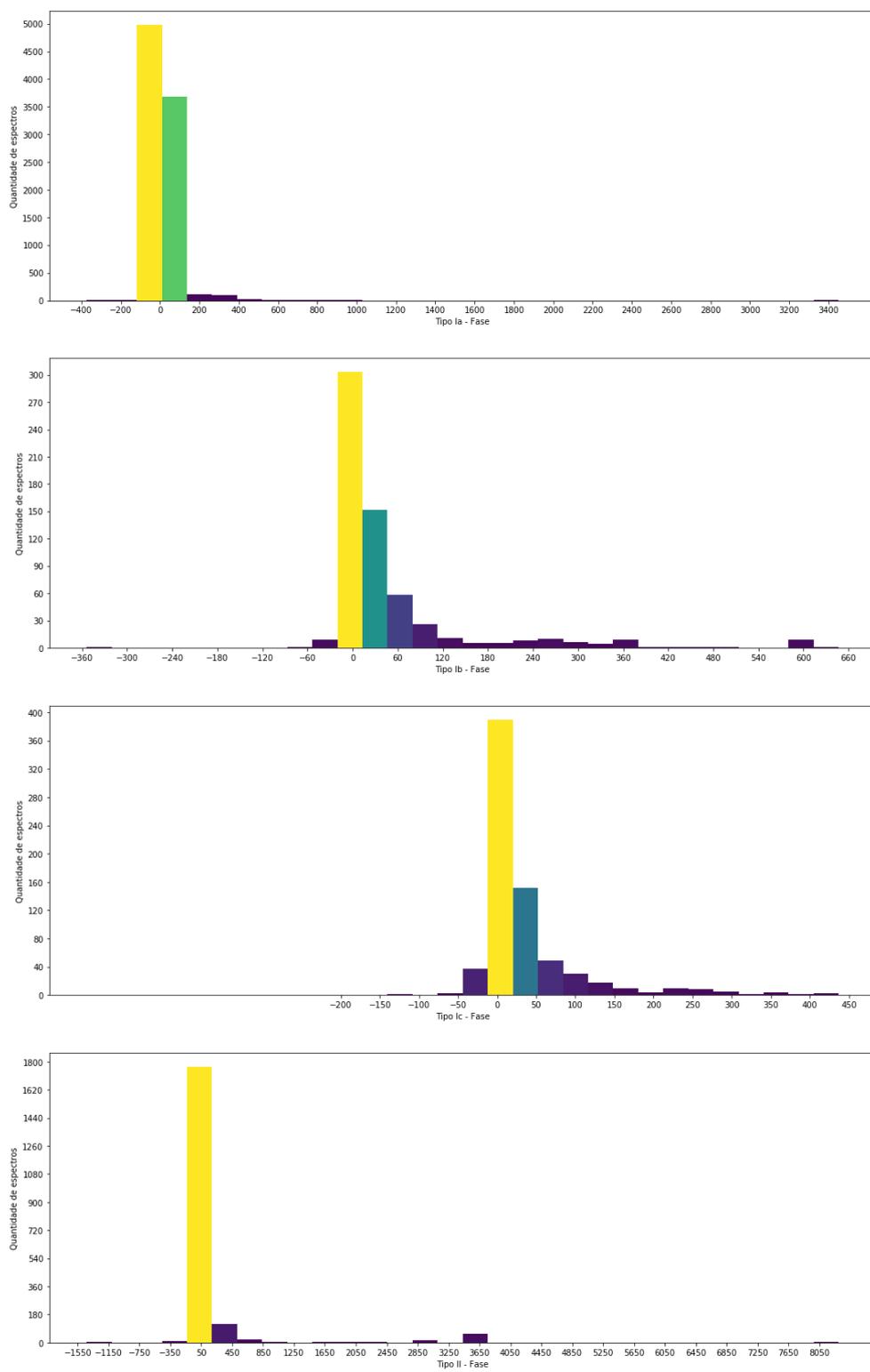
$$Y_j = \frac{\sum_{i=-m}^m C_i Y_{j+i}}{N} \quad (5.2)$$

Onde:

- Y_j : resultado da suavização;
- C_i : coeficiente da i -ésima suavização;
- N : tamanho da janela de pontos do filtro (equivalente a $2m+1$).

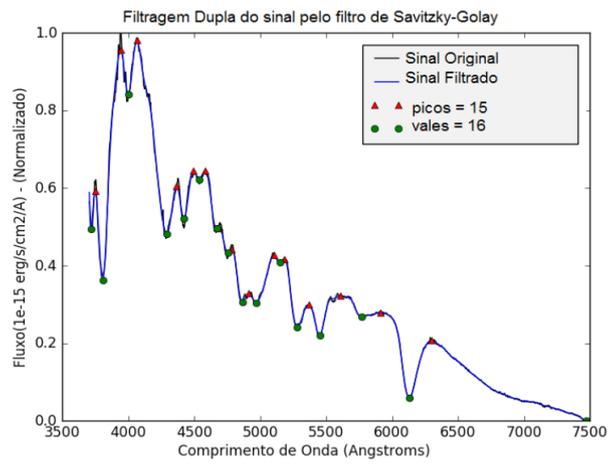
Os espectros passam pelo filtro de SG duas vezes consecutivas. Os parâmetros usados em cada filtragem são: tamanho da janela de pontos = 71 e grau do polinômio = 9, ambos os parâmetros são os mesmos definidos por (ARANTES FILHO, 2018). A Figura 5.3 mostra o resultado da Dupla-Filtragem em comparação com o sinal original de um espectro, a preservação dos picos e vales é destacada.

Figura 5.2 - Histogramas de fase espectral por tipo de SNs.



Fonte: Produção da autora.

Figura 5.3 - Resultado da dupla filtragem de um espectro.



Fonte: Arantes Filho (2018).

6 MÉTRICAS DE AVALIAÇÃO DO DESEMPENHO

Algumas métricas de avaliação do desempenho da classificação por aprendizagem de máquina são adotadas neste trabalho. As fórmulas são as mesmas apresentadas por (SOKOLOVA; LAPALME, 2009) para classificadores binários e classificação multi-classe. Todas elas usam as informações da matriz de confusão, que é uma tabela que mostra os acertos e os erros da classificação (Tabela 6.1). Adotar métricas bem definidas é importante para que a avaliação do desempenho seja feita de forma objetiva e os resultados possam ser usados para comparações com trabalhos posteriores, tanto de otimização como de criação de novas metodologias.

Tabela 6.1 - Modelo de uma matriz de confusão.

Classificação do dado	Positivo	Negativo
Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Sokolova e Lapalme (2009).

As métricas de classificação do desempenho para classificadores binários, derivadas da matriz de confusão são as seguintes:

- Acurácia: indica o quão frequente o classificador acerta;

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN} \quad (6.1)$$

- Precisão: indica a concordância entre os positivos encontrados pelo classificador e os que são realmente positivos;

$$Precisão = \frac{VP}{VP + FP} \quad (6.2)$$

- *Recall* ou Sensibilidade: indica a eficácia para encontrar os positivos;

$$Recall = \frac{VP}{VP + FN} \quad (6.3)$$

- *F1-Score*: relaciona Precisão e *Recall* para indicar a qualidade do classifi-

gador;

$$F1 - Score = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad (6.4)$$

- Especificidade: indica o quão efetivo é o classificador para identificar os negativos;

$$Especificidade = \frac{VN}{FP + VN} \quad (6.5)$$

- AUC (do inglês, *Area Under the Curve*): mede a capacidade do classificador de evitar as classificações falsas.

$$AUC = \frac{Recall + Especificidade}{2} \quad (6.6)$$

Além das medidas descritas acima, também foi adotado o índice Kappa que mede a concordância entre duas ou mais classificações. Neste trabalho, compara-se a classificação feita pelas RNAs que compõem a CINTIA 2 com a classificação feita pelos especialistas humanos. A Equação 6.7 mostra como calcular o índice Kappa, de acordo com (LANDIS; KOCH, 1977), que resulta na interpretação mostrada na Tabela 6.2.

$$k = \frac{\pi_0 - \pi_e}{1 - \pi_e} \quad (6.7)$$

Onde:

- π_0 : observado;
- π_e : esperado.

Tabela 6.2 - Interpretação do índice Kappa.

Índice Kappa	Força da Concordância
< 0.00	Pobre
0.00 - 0.20	Fraca
0.21 - 0.40	Razoável
0.41 - 0.60	Moderada
0.61 - 0.80	Substancial
0.81 - 1.00	Quase Perfeita

Fonte: Adaptado de Landis e Koch (1977).

A partir da matriz de confusão e das métricas de classificadores binários, depreende-se as métricas de avaliação para a classificação multi-classe, que se mostram importantes para medir o desempenho da CINTIA 2. As equações para calcular as métricas também são relatadas por (SOKOLOVA; LAPALME, 2009). Para os problemas de multi-classe, há dois grupos principais de métricas, as macro (M) e as micro (μ). As métricas M são boas para pontuar a contribuição média de cada classe, enquanto as μ agregam o peso das contribuições individuais. As métricas μ exercem papel importante na avaliação de classificadores desbalanceados, quando a quantidade de padrões de cada classe é bastante díspar uma da outra, que é o caso da CINTIA 2 (como pode ser visto na Figura 5.1). As métricas são descritas em seguida.

- Acurácia Média: eficácia média de acertos por classe; onde k=número de classes.

$$AcuráciaMédia = \frac{\sum_{i=1}^k Acurácia}{k} \quad (6.8)$$

- Taxa de Erro: média do erro de classificação por classe; onde k=número de classes.

$$TaxadeErro = 1 - AcuráciaMédia \quad (6.9)$$

- Precisão M: precisão macro é a média das precisões individuais; onde k=número de classes.

$$PrecisãoM = \frac{\sum_{i=1}^k Precisão}{k} \quad (6.10)$$

- Recall M: Recall macro é a média dos recalls individuais; onde k=número de classes.

$$RecallM = \frac{\sum_{i=1}^k Recall}{k} \quad (6.11)$$

- F1-Score M: F1-Score macro é a média dos F1-Score individuais; onde k=número de classes.

$$F1 - ScoreM = \frac{\sum_{i=1}^k F1 - Score}{k} \quad (6.12)$$

- Precisão μ : Precisão micro é a Precisão média do classificador, tomando os valores individuais da matriz de confusão de cada classe; onde k=número de classes.

$$Precisão\mu = \frac{\sum_{i=1}^k VP_i}{\sum_{i=1}^k VP_i + FP_i} \quad (6.13)$$

- *Recall μ* : *Recall* micro é o *Recall* médio do classificador, tomando os valores individuais da matriz de confusão de cada classe; onde k =número de classes.

$$Recall\mu = \frac{\sum_{i=1}^k VP_i}{\sum_{i=1}^k VP_i + FN_i} \quad (6.14)$$

- *F1-Score μ* : *F1-Score* micro é o *F1-Score* médio do classificador, tomando os valores individuais da matriz de confusão de cada classe; onde k =número de classes.

$$F1 - Score\mu = \frac{2 * Precisão\mu * Recall\mu}{Precisão\mu + Recall\mu} \quad (6.15)$$

7 DESENVOLVIMENTO DA CINTIA 2

Aborda-se neste capítulo o processo de desenvolvimento da CINTIA 2, desde a extração das entradas, as configurações testadas para cada RNA e os resultados alcançados em cada RNA individual que levaram ao desenho da arquitetura integradora das redes, que está sendo proposta neste trabalho e foi apresentada preliminarmente por (NASCIMENTO et al., 2019). A concepção de uma arquitetura que integra as RNAs, proporcionando classificações sem risco de ambiguidade, é uma das melhorias aplicada na CIntIa, transformando-lhe assim em CINTIA 2.

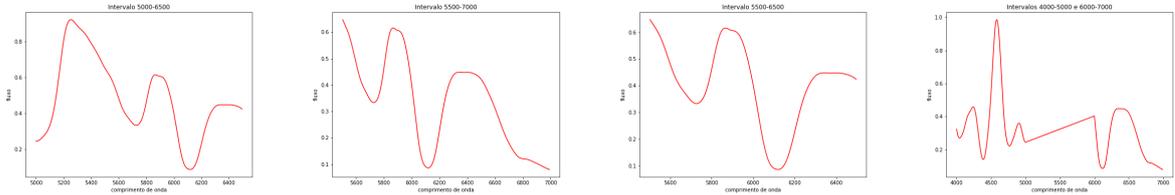
7.1 Extração das Entradas

A extração das entradas é a etapa que vem em seguida ao pré-processamento dos espectros, as características necessárias para o treinamento das RNAs que compõem o sistema são extraídas dos espectros. Primeiramente, cada espectro passa por uma interpolação a cada 8 pontos, começando em 4000 Å até 7000 Å, que resulta em 375 pontos. As entradas de cada RNA são um conjunto de valores de fluxo correspondentes a posições do comprimento de onda, contidos em intervalos definidos originalmente por (M6DOLO, 2016), que se referem às regiões em que os elementos H, Si, S e He se manifestam e são as mesmas regiões analisadas por especialistas humanos. Os intervalos são os seguintes:

- 5000 a 6500Å: para classificar em tipo Ia ou não;
- 5500 a 7000Å: para classificar em tipo Ib ou não;
- 5500 a 6500Å: para classificar em tipo Ic ou não;
- 4000 a 5000Å e 6000 a 7000Å: para classificar em tipo II ou não.

Cada espectro origina quatro conjuntos de entrada, um para cada RNA binária. A Figura 7.1 mostra uma representação gráfica desses conjuntos para um espectro da SN2002bo.

Figura 7.1 - Representação gráfica dos intervalos escolhidos como entradas das 4 RNAs. Começando da esquerda os intervalos são referentes às RNAs: Ia, Ib, Ic e II.



Fonte: Produção da autora.

7.2 Parâmetros de Treinamento

Alguns parâmetros são necessários para executar o treinamento de uma RNA. As taxas de aprendizado e de *momentum* são responsáveis pela estabilidade do algoritmo retropropagação. O *bias* é um fator que serve para diminuir ou aumentar a saída do somador de cada neurônio. A função de ativação é a função que define a saída de cada neurônio. Além disso, é preciso definir uma taxa de erro que será tolerado pelo algoritmo e os pesos iniciais. A quantidade máxima de épocas executadas foi estabelecida como 100000 (cem mil), logo se o algoritmo não convergir para o erro tolerado em até 100000 iterações o treinamento é encerrado. Todos esses parâmetros, exceto pela quantidade máxima de épocas, foram definidos por (MóDOLO, 2016) quando desenvolveu a primeira versão da CIntIa e, portanto replicados pela CINTIA 2. Os parâmetros são discriminados abaixo:

- Taxa de aprendizagem = 0.5;
- *Momentum* = 0.3;
- *Bias* = 1;
- Erro tolerado = 0.001;
- Função de ativação = sigmóide;
- Pesos iniciais: escolhidos aleatoriamente.

7.3 Topologias Testadas

Foram executados 7 treinamentos para cada RNA, como mostrado na Tabela 7.1, a fim de escolher a topologia mais adequada para cada uma delas. A escolha das topologias utilizadas nos treinamentos segue os experimentos realizados para a CIntIa. (MÓDOLO, 2016) executou 8 topologias durante os treinamentos de cada RNA, neste trabalho, optou-se por replicar os treinamentos de 6 topologias (2 a 7 na Tabela 7.1) mais a topologia 1 (Tabela 7.1) cuja estrutura em termos de camadas e neurônios é a mais simples. O máximo de camadas ocultas permitido pelo sistema é duas. Os parâmetros e topologias foram testados manualmente. Em cada treinamento foram usados 60% dos dados do conjunto, enquanto foram reservados 40% para testes.

Tabela 7.1 - Número de neurônios em cada topologia testada.

Id. da Topologia	Neurônios na camada 1	Neurônios na camada 2
1	5	0
2	10	0
3	25	0
4	40	0
5	10	10
6	20	4
7	40	8

Fonte: Produção da autora.

7.4 Treinamento e Teste da RNA Ia

Foram executados 7 treinamentos da RNA binária que classifica entre espectros de SNs do tipo Ia e espectros de SNs que não são do tipo Ia (tipo Ia ou tipo não-Ia). Cada treinamento passou por duas rodadas de testes, em cada teste foram utilizados 20% dos dados disponíveis.

Os espectros do tipo Ia apresentam um padrão muito característico principalmente na fase espectral de brilho máximo, que para (BLONDIN et al., 2011) está situada entre -3 e +3 dias. Outros autores expandem um pouco esses limites, em torno de -5 e +5 dias ainda é possível ver claramente o padrão esperado na maioria dos espectros de SNs Ia. Esse padrão diz respeito às linhas espectrais de Si e S que tem presença marcante nas SNs Ia. Por isso, experimentamos expandir ainda mais os limites com a intenção de deixar o classificador o mais geral possível. O alargamento

da amplitude das fases espectrais aceitas foi feito de forma incremental e em bloco, criamos blocos de fases (Tabela 7.2) e adicionamos um novo bloco a cada sessão de treinamento, iniciando pelo bloco de Máximo. Estabelecemos o piso para Acurácia de 0.90 (90%) e índice Kappa 0.81 (Quase Perfeito), mínimos exigidos para atestar a qualidade da classificação.

Tabela 7.2 - Bloco de fases espectrais definidas para auxiliar na diversificação dos espectros de SNIa treinados.

Id. do Bloco	Nome do bloco	Amplitude de fases espectrais
1	Pré-Pré-Máximo	≤ -11 dias
2	Pré-Máximo	> -11 e ≤ -4 dias
3	Máximo	> -4 e $< +4$ dias
4	Pós-Máximo	$\geq +4$ e $< +11$ dias
5	Pré-Nebular	$\geq +11$ e $< +46$ dias
6	Nebular	$\geq +46$ dias

Fonte: Produção da autora.

Uma sessão de treinamento consiste em executar 7 experimentos, um para cada topologia da Tabela 7.1, com um bloco de espectros a mais, a partir da 2ª sessão. Foram executadas 5 sessões de treinamento porque a partir da inclusão dos blocos Pré-Pré-Máximo e Pré-Nebular as métricas de referência (Acurácia e Índice Kappa) ficaram abaixo do piso. Os melhores resultados de cada sessão de treinamentos são apresentados na Tabela 7.3, que mostra os resultados das métricas de referência e o Id. da topologia definido na Tabela 7.1.

Tabela 7.3 - Melhores resultados das sessões de treinamento realizadas para definir a amplitude das fases espectrais dos espectros do tipo Ia incluídos no classificador.

Id. da sessão	Id. da topologia	Blocos de fases	Acurácia	Kappa
1	2	3	0.9530	0.8978
2	1	3 e 4	0.9223	0.8464
3	1	2, 3 e 4	0.9196	0.8385
4	1	2, 3, 4 e 5	0.7079	0.4561
5	3	1, 2, 3 e 4	0.8905	0.7809

Fonte: Produção da autora.

O critério de escolha da amplitude de fases espectrais adotado é o melhor resultado de teste combinado com a maior quantidade de blocos de fases espectrais, observando os valores de piso das métricas de referência. Portanto, consideramos que a classificação mais adequada para o nosso propósito foi a da sessão 3 (Tabela 7.3), onde as métricas de referência estão acima do piso e englobam espectros entre as fases > -11 e $< +11$ dias (Pré-Máximo, Máximo e Pós-Máximo). Assim, aumentou-se a amplitude de fases dos espectros das SNs Ia, em comparação com todos os trabalhos consultados, o que contribui para a diversificação dos padrões aprendidos pela RNA.

7.4.1 Detalhamento do Melhor Resultado

O melhor resultado, alcançado como descrito anteriormente, foi com a topologia 1: uma camada com 5 neurônios. Vale ressaltar que os espectros de SNs Ib, Ic e II, que compõem a classe tipo não-Ia, não tem limitação de fase espectral (todas as fases estão incluídas). E os espectros de SNs Ia estão nas fases espectrais > -11 e $< +11$.

Foram realizados testes com dois conjuntos para avaliar a consistência da classificação efetuada pela RNA. A Tabela 7.4 apresenta uma matriz de confusão adaptada para mostrar os valores dos dois testes. Enquanto, a Tabela 7.5 exhibe as métricas de avaliação do desempenho para os dois testes. Observa-se que ambos os testes alcançaram resultados similares, o Teste 2 é ligeiramente melhor do que o Teste 1, assim os seus resultados foram usados para calcular as métricas de avaliação para a classificação multi-classe, o que será abordado posteriormente.

Tabela 7.4 - Matriz de confusão da RNA Ia adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	609	94	6	486
Teste 2	609	93	3	489

Fonte: Produção da autora.

7.5 Treinamento e Teste da RNA Ib

Foram executados 7 treinamentos da RNA binária que classifica entre espectros de SNs do tipo Ib e espectros de SNs que não são do tipo Ib (tipo Ib ou tipo não-Ib). Cada treinamento passou por duas rodadas de testes, em cada teste foram utilizados 20% dos dados disponíveis. Diferentemente dos espectros de SNs Ia, os espectros de SNs Ib não tem padrões muito bem definidos, apesar de ser consenso que a presença

Tabela 7.5 - Métricas de avaliação do desempenho da RNA Ia dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9163	0.9196
Precisão	0.9902	0.9951
Recall	0.8663	0.8675
F1-Score	0.9241	0.9269
Especificidade	0.9878	0.9939
AUC	0.9270	0.9307
Índice Kappa	0.8318	0.8385

Fonte: Produção da autora.

de linhas de He e ausência de linhas de H são as características fundamentais usadas na sua classificação.

Realizou-se, primeiramente, treinamentos usando restrição de fases espectrais para os espectros do tipo Ib, apenas espectros de SNs Ib com fases entre -4 e +4 dias (brilho máximo) foram incluídos. Enquanto os espectros dos outros tipos foram incluídos sem restrições de fase espectral. O melhor resultado obtido foi com uma camada e 5 neurônios, a Tabela 7.6 apresenta a matriz de confusão e a Tabela 7.7 as métricas de avaliação dos dois testes realizados para esse treinamento.

Tabela 7.6 - Matriz de confusão da RNA Ib com espectros Ib, apenas na fase de brilho máximo, adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	10	3	3	1692
Teste 2	7	6	5	1690

Fonte: Produção da autora.

Em seguida, realizou-se novos treinamentos sem aplicar a restrição de fases espectrais para os espectros de SNs Ib. Dessa vez, o melhor resultado alcançado foi com a topologia 6: duas camadas com 20 neurônios na 1ª camada e 4 neurônios na 2ª. Foram realizados testes com dois conjuntos para avaliar a consistência da classificação efetuada pela RNA. A Tabela 7.8 apresenta uma matriz de confusão adaptada para mostrar os valores dos 2 testes. Enquanto, a Tabela 7.9 exhibe as métricas de avaliação do desempenho para os 2 testes. Percebe-se que a diferença do melhor resultado quando se aplica a restrição de fases e quando não se aplica é pequena,

Tabela 7.7 - Métricas de avaliação do desempenho da RNA Ib, apenas na fase de brilho máximo, dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9965	0.9936
Precisão	0.7692	0.5833
Recall	0.7692	0.5385
F1-Score	0.7692	0.5600
Especificidade	0.9982	0.9971
AUC	0.8837	0.7678
Índice Kappa	0.7675	0.5568

Fonte: Produção da autora.

considerando-se a interpretação do índice Kappa, ambos os resultados tem concordância Substancial. Outro fator considerado é a quantidade de espectros testados, com a restrição de fases foram testados apenas 26 espectros de SNs Ib e sem a restrição foram 178. Sendo assim, optou-se por usar todos os espectros de SNs Ib na classificação, sem as restrições de fase espectral. Os espectros de SNs Ia, Ic e II, que compõem a classe tipo não-Ib, também não tem limitação de fase espectral.

Tabela 7.8 - Matriz de confusão da RNA Ib adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	52	37	8	1687
Teste 2	42	47	11	1684

Fonte: Produção da autora.

Tabela 7.9 - Métricas de avaliação do desempenho da RNA Ib dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9748	0.9675
Precisão	0.8667	0.7925
Recall	0.5843	0.4719
F1-Score	0.96980	0.5915
Especificidade	0.9953	0.9935
AUC	0.7898	0.7327
Índice Kappa	0.6853	0.5758

Fonte: Produção da autora.

Observa-se que os 2 testes do melhores treinamentos, com e sem restrição de fases, alcançaram resultados diferentes. No caso dos testes sem restrição de fases (Tabela 7.9), o Teste 1 é melhor do que o Teste 2, expresso no Índice Kappa com uma diferença de mais de 0.10 pontos. A diferença observada é uma evidência do pressuposto adotado no início desta seção, quando foi admitido que os padrões dos espectros de SNs Ib não são bem definidos, podendo existir mais de um padrão clássico. Ainda assim foram alcançados resultados melhores do que a 1ª versão da CIntIa e com uma quantidade muito maior de padrões de teste o que permite fazer uma análise mais confiante. Adotou-se o melhor resultado (Teste 1) para fins de cálculo das métricas de avaliação do classificador multi-classe, que será feito posteriormente.

7.6 Treinamento e Teste da RNA Ic

Foram executados 7 treinamentos da RNA binária que classifica entre espectros de SNs do tipo Ic e espectros de SNs que não são do tipo Ic (tipo Ic ou tipo não-Ic). Cada treinamento passou por duas rodadas de testes, em cada teste foram utilizados 20% dos dados disponíveis. Assim como os espectros de SNIb, os espectros de SNIc não tem padrões muito bem definidos e podem até ser confundidos, em determinadas regiões, com espectros de SNs Ib devido a origem comum dos dois.

Realizou-se, primeiramente, treinamentos usando restrição de fases espectrais para os espectros do tipo Ic, apenas espectros de SNs Ic com fases entre -4 e +4 dias (brilho máximo) foram incluídos. Enquanto os espectros dos outros tipos foram incluídos sem restrições de fase espectral. O melhor resultado obtido foi com uma camada e 25 neurônios, a Tabela 7.10 apresenta a matriz de confusão e a Tabela 7.11 as métricas de avaliação dos dois testes realizados para esse treinamento.

Tabela 7.10 - Matriz de confusão da RNA Ic com espectros Ic, apenas na fase de brilho máximo, adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	4	11	8	1668
Teste 2	5	10	4	1672

Fonte: Produção da autora.

Tabela 7.11 - Métricas de avaliação do desempenho da RNA Ic, apenas na fase de brilho máximo, dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9888	0.9917
Precisão	0.333	0.5556
Recall	0.2667	0.3333
F1-Score	0.2963	0.4167
Especificidade	0.9952	0.9976
AUC	0.6309	0.6655
Índice Kappa	0.2907	0.4128

Fonte: Produção da autora.

Em seguida, realizou-se novos treinamentos sem aplicar a restrição de fases espectrais para os espectros de SNs Ic. O melhor resultado alcançado também foi com a topologia 3: uma camada com 25 neurônios. Foram realizados testes com dois conjuntos para avaliar a consistência da classificação efetuada pela RNA. A Tabela 7.12 apresenta uma matriz de confusão adaptada para mostrar os valores dos dois testes. Enquanto, a Tabela 7.13 exibe as métricas de avaliação do desempenho para os dois testes. Percebe-se que existe uma diferença a ser considerada entre o melhor resultado quando se aplica a restrição de fases (concordância Moderada) e quando não se aplica (concordância Razoável). Porém, com a restrição de fases apenas 30 espectros de SNs Ic foram testados, enquanto sem a restrição de fases foram testados 212. Sendo assim, optou-se por usar todos os espectros de SNs Ic na classificação, sem a restrições de fase espectral. Os espectros de SNs Ia, Ib e II, que compõem a classe tipo não-Ic, também não tem limitação de fase espectral.

Tabela 7.12 - Matriz de confusão da RNA Ic adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	20	86	9	1667
Teste 2	23	83	4	1672

Fonte: Produção da autora.

Tabela 7.13 - Métricas de avaliação do desempenho da RNA Ic dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9467	0.9512
Precisão	0.6897	0.8519
Recall	0.1887	0.2170
F1-Score	0.2963	0.3459
Especificidade	0.9946	0.9976
AUC	0.5917	0.6073
Índice Kappa	0.2779	0.3297

Fonte: Produção da autora.

No caso dos testes sem restrição de fase, observamos que ambos alcançaram resultados apenas ligeiramente diferentes, o Teste 2 é apenas um pouco melhor do que o Teste 1. Os resultados não são bons, a concordância indicada pelo Índice Kappa é Razoável o que indica que os padrões não são satisfatoriamente reconhecidos ou que não existe um padrão clássico que contemple a maioria dos espectros de SNS Ic conhecidos. Adotou-se o melhor resultado (Teste 2) para fins de cálculo das métricas de avaliação do classificador multi-classe, que será feito posteriormente.

7.7 Treinamento e Teste da RNA II

Foram executados 7 treinamentos da RNA binária que classifica entre espectros de SNs do tipo II e espectros de SNs que não são do tipo II (tipo II ou tipo não-II). Cada treinamento passou por duas rodadas de testes, em cada teste foram utilizados 20% dos dados disponíveis.

Realizou-se, primeiramente, treinamentos usando restrição de fases espectrais para os espectros do tipo II, apenas espectros de SNs II com fases entre -4 e +4 dias (brilho máximo) foram incluídos. Enquanto os espectros dos outros tipos foram incluídos sem restrições de fase espectral. O melhor resultado obtido foi com duas camadas com 10 neurônios cada, a Tabela 7.14 apresenta a matriz de confusão e a Tabela 7.15 as métricas de avaliação dos dois testes realizados para esse treinamento.

Tabela 7.14 - Matriz de confusão da RNA II com espectros II, apenas na fase de brilho máximo, adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	44	8	6	1483
Teste 2	43	10	17	1472

Fonte: Produção da autora.

Tabela 7.15 - Métricas de avaliação do desempenho da RNA II, apenas na fase de brilho máximo, dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9909	0.9825
Precisão	0.8800	0.7167
Recall	0.8462	0.8113
F1-Score	0.8627	0.7611
Especificidade	0.9960	0.9886
AUC	0.9211	0.8999
Índice Kappa	0.8580	0.7520

Fonte: Produção da autora.

Em seguida, realizou-se novos treinamentos sem aplicar a restrição de fases espectrais para os espectros de SNs II. O melhor resultado alcançado foi com a topologia 6: duas camadas com 20 neurônios na 1ª camada e 4 na 2ª camada. Foram realizados testes com 2 conjuntos para avaliar a consistência da classificação efetuada pela RNA. A Tabela 7.16 apresenta uma matriz de confusão adaptada para mostrar os valores dos testes. Enquanto, a Tabela 7.17 exhibe as métricas de avaliação do desempenho para os dois testes. Observa-se que os dois alcançaram resultados similares, o Teste 1 é um pouco melhor do que o Teste 2. Os resultados são muito bons, superiores aos alcançados pela CIntIa.

Tabela 7.16 - Matriz de confusão da RNA II adaptada para os 2 testes.

	VP	FN	FP	VN
Teste 1	230	67	8	1481
Teste 2	223	74	16	1473

Fonte: Produção da autora.

Tabela 7.17 - Métricas de avaliação do desempenho da RNA II dos 2 testes.

	Teste 1	Teste 2
Acurácia	0.9580	0.9496
Precisão	0.9664	0.9331
Recall	0.7744	0.7508
F1-Score	0.8598	0.8321
Especificidade	0.9946	0.9893
AUC	0.8845	0.8700
Índice Kappa	0.8355	0.8029

Fonte: Produção da autora.

Percebe-se que a diferença do melhor resultado quando se aplica a restrição de fases e quando não se aplica é pequena, considerando-se a interpretação do índice Kappa, ambos os resultados tem concordância Quase Perfeita. Outro fator considerado é a quantidade de espectros testados, com a restrição de fases foram testados 105 espectros de SNs II e sem a restrição foram 594. Sendo assim, optou-se por usar todos os espectros de SNs II na classificação, sem a restrições de fase espectral. Os espectros de SNs Ia, Ib e Ic, que compõem a classe tipo não-II, também não tem limitação de fase espectral. Assim, adotou-se o melhor resultado (Teste 1) sem a restrição de fases para fins de cálculo das métricas de avaliação do classificador multi-classe, que será feito posteriormente.

7.8 Arquitetura da CINTIA 2

A arquitetura proposta resulta dos testes realizados em cada uma das RNAs binárias após os treinamentos. Uma maneira de integrar as RNAs individuais não foi proposta no desenvolvimento da 1ª versão da CIntIa, assim havia a possibilidade do classificador produzir respostas ambíguas. A arquitetura apresentada aqui foi

desenhada a fim de dirimir esse problema.

Cada uma das RNAs é um módulo na visão geral da arquitetura, a Figura 7.2 mostra as topologias das RNAs em cada módulo. Enquanto a Figura 7.3 apresenta a visão geral da arquitetura da CINTIA 2. O fluxo de classificação passa por uma hierarquia de RNAs, que é uma forma de aplicar o ranking de classes (fase de reconstrução) definido no Capítulo 3. A ordem dos módulos na hierarquia foi definida seguindo os valores de Índice Kappa, o módulo com maior índice é ativado primeiro, o critério de desempate é o valor do *F1-Score*.

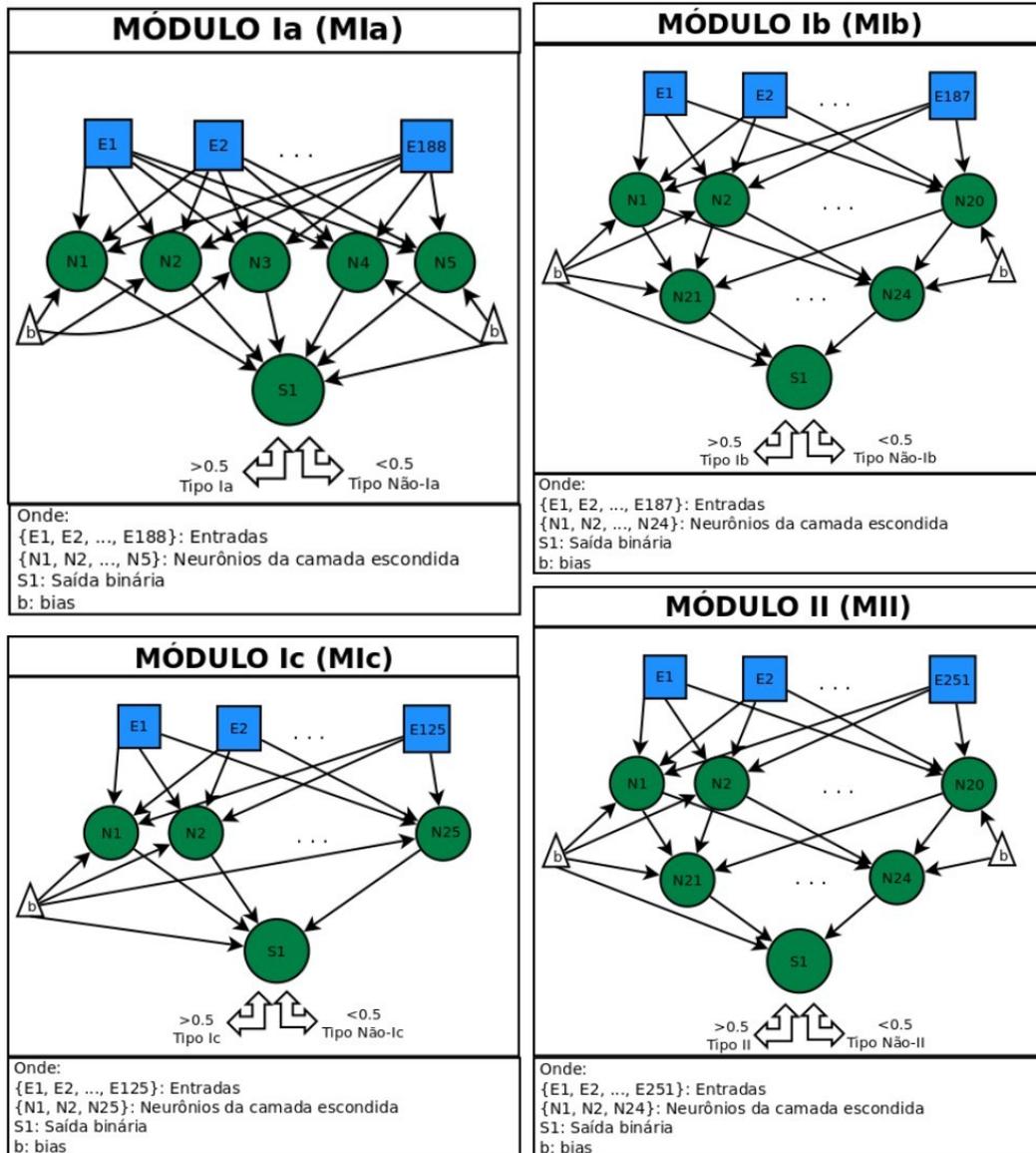
O fluxo da classificação inicia identificando se o espectro, previamente pré-processado, é do tipo Ia ou não. Em seguida, são classificados os do tipo II, único tipo que apresenta linhas espectrais de H. Na sequência os espectros devem ser classificados como tipo Ib ou tipo Ic, que é a diferenciação mais desafiadora a ser feita pelo sistema inteligente, assim como é para os especialistas humanos. Caso o espectro não seja classificado em nenhum dos módulos ele recebe o rótulo de “Tipo Não Identificado”. A Tabela 7.18 mostra as métricas de avaliação da CINTIA 2 considerando o sistema como um classificador multi-classe, foram usados os melhores resultados de cada RNA para calcular as métricas do sistema integrado. As métricas M são mais baixas das que as métricas μ porque elas não ponderam as contribuições, mas aplicam uma média simples. As métricas μ dão mais peso às classes com maior quantidade de padrões, no caso tipo Ia e tipo II, que também são as que apresentam melhores índices individuais. Assim as nuances do classificador são observadas e o desempenho medido é significativamente melhor.

Tabela 7.18 - Métricas de avaliação do desempenho da CINTIA 2.

Acurácia Média	0.9509
Taxa de Erro	0.0491
Precisão M	0.9200
Recall M	0.6108
F1-Score M	0.7077
Precisão μ	0.9755
Recall μ	0.7655
F1-Score μ	0.8624

Fonte: Produção da autora.

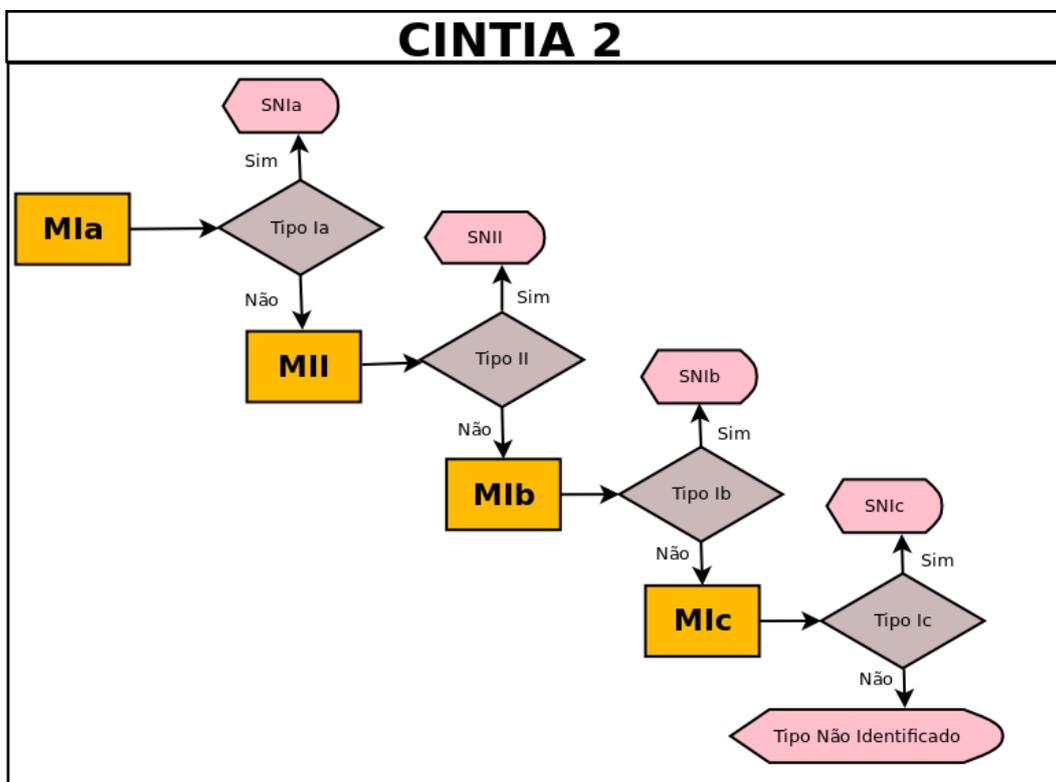
Figura 7.2 - Módulos que compõem a CINTIA 2.



Os quadrados azuis representam as entradas, os círculos verdes são os neurônios e os triângulos brancos os bias.

Fonte: Produção da autora.

Figura 7.3 - Arquitetura da CINTIA 2 desenhada como um fluxograma.



Fonte: Produção da autora.

8 DESENVOLVIMENTO DO SOFTWARE CINTIA 2

O *software* CINTIA 2 é a implementação da arquitetura desenhada e apresentada no capítulo anterior. Foram usadas as linguagens de programação C++ e Python. Python foi escolhida por ser uma linguagem de fácil aprendizagem e que dispõe de diversas bibliotecas de aprendizagem de máquina e manipulação de dados, por isso é uma das linguagens mais utilizadas atualmente por quem programa sistemas de Inteligência Artificial. Utilizou-se a linguagem C++ para reaproveitar componentes que haviam sido programados para a CIntIa, mais precisamente, as RNAs (treinamento e ativação). Portanto, o software produzido integra código das duas linguagens de programação, ambas de código aberto e portáteis entre sistemas operacionais diferentes.

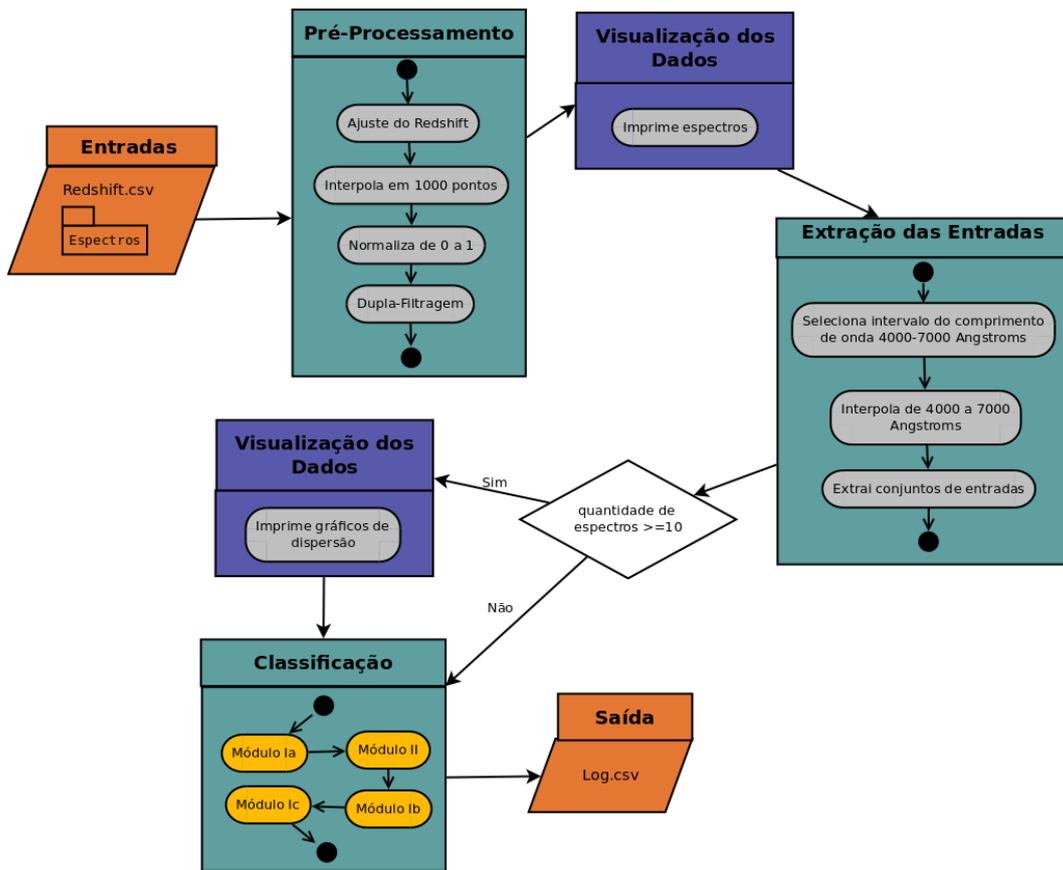
Usou-se a versão 3 do Python, que pode conter incompatibilidades com versões anteriores, por isso o bom funcionamento é garantido apenas para computadores que tenham o Python 3 instalado. Algumas bibliotecas de Python foram utilizadas, as quais são todas livres para uso e descritas em seguida:

- csv: módulo do Python usado para leitura e escrita em arquivos *Comma Separated Values*, de extensão .csv;
- datetime: módulo do Python para manipular data e hora;
- matplotlib: biblioteca usada para imprimir gráficos;
- numpy: biblioteca de computação científica, que permite manipular vetores e diversas funções matemáticas;
- os: módulo do Python para acessar e fazer operações no sistema operacional;
- pandas: biblioteca usada para análise de dados e uma das mais importantes para implementar soluções de ciência de dados;
- scipy: ecossistema de bibliotecas de uso científico e engenharia;
- sklearn: biblioteca que integra algoritmos de aprendizagem de máquina.

O *software* foi desenvolvido para incorporar as etapas de pré-processamento até a classificação dos espectros. A automatização das etapas também é um facilitador para utilização da metodologia de classificação apresentada neste trabalho.

Elaborou-se um fluxograma (Figura 8.1) que mostra todo o percurso de processamento do espectro. Explica-se com mais detalhes os componentes e processos da CINTIA 2 nas seções posteriores.

Figura 8.1 - Fluxograma que representa o funcionamento do software CINTIA 2.



Fonte: Produção da autora.

8.1 Entrada e Saída

Os valores de *redshift* e os arquivos contendo os espectros a serem classificados são as entradas exigidas pela CINTIA 2. Os *redshifts* devem constar em único arquivo de extensão .csv com pelo menos duas colunas, uma coluna com o nome da SN e outra com o próprio valor do *redshift*. A outra entrada é uma pasta com os espectros (um ou mais) que serão classificados. Cada espectro é um arquivo, também com extensão .csv, com duas colunas, a primeira coluna contém os comprimentos de onda e a segunda, os fluxos correspondentes. Em virtude da interação homem-máquina

proporcionada pelo software ser por meio de um *prompt* de comando, as entradas são os caminhos (formato texto) até o local em que os arquivos estão armazenados.

A saída, resultado da classificação efetuada pela hierarquia de RNAs binárias, é apresentada ao usuário de duas formas. A primeira forma é a exibição no *prompt* de comando (Figura 8.2) e a segunda é a criação de um arquivo .csv com o nome das SNs, a identificação dos seus espectros e a classificação, ou seja, o tipo de SN a que pertencem. O arquivo criado é salvo em uma pasta chamada "Logs", que pode ser consultado posteriormente e facilita o envio para outras pessoas interessadas.

Figura 8.2 - Tela que apresenta os resultados da classificação feita pela CINTIA 2.

```
Generating files...
Classifying...
Running module II...
Running module Ia...
Running module Ib...
Running module Ic...

-----CLASSIFICATION-----
Pattern: 1
Supernova: SN1994D
Spectrum: SN1994D-1
Classification: Tipo Ia

Pattern: 2
Supernova: SN1994D
Spectrum: SN1994D-10
Classification: Tipo Ia

Pattern: 3
Supernova: SN1994D
Spectrum: SN1994D-11
Classification: Tipo Ia

Pattern: 4
Supernova: SN1994D
Spectrum: SN1994D-12
Classification: Tipo Ia

Pattern: 5
Supernova: SN1994D
Spectrum: SN1994D-13
Classification: Tipo Ia
```

Fonte: Produção da autora.

8.2 Processamento

O processamento funciona em três blocos de atividades. O primeiro bloco, o de pré-processamento, é acionado logo que as entradas são fornecidas pelo usuário do sistema, é nesta etapa que todas as ações descritas na Seção 5.2 são executadas. O ajuste do *redshift* é o primeiro processo seguido pela interpolação em 1000 pontos

e normalização de 0 a 1, passos importantes para que a Dupla-Filtragem funcione corretamente.

O bloco de extração das entradas seleciona os espectros que têm o intervalo de comprimento de onda na faixa do visível, apenas estes podem ser classificados pela CINTIA 2. Os espectros que continuam no conjunto após a seleção são interpolados entre 4000 e 7000 Å, de forma que todos eles fiquem com a mesma quantidade de pontos. Então, os conjuntos de entradas, cada espectro gera quatro conjuntos de pontos, são preparados para serem usados na etapa de classificação, que é o último bloco de processamento representado na Figura 8.1.

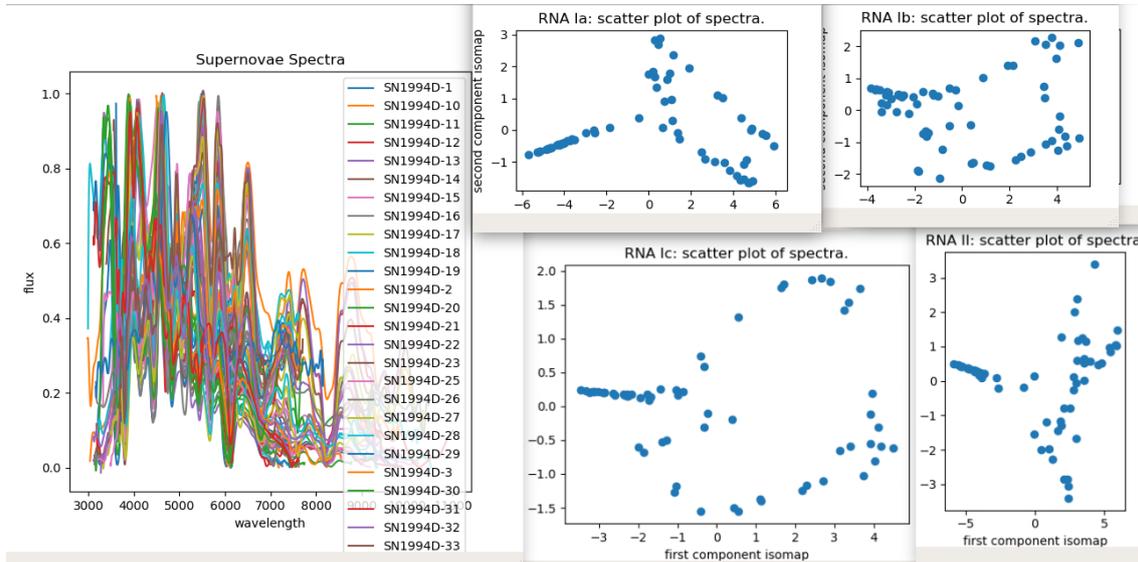
A classificação, que já foi detalhada quando foi abordada a arquitetura da CINTIA 2 (Capítulo 7), agrupa os quatro módulos ativados de acordo com a hierarquia proposta neste trabalho. Em termos práticos, cada módulo é um arquivo executável gerado pelo código C++ que além de fazer o treinamento da RNA também delinea a sua ativação quando as RNAs já estão treinadas, ou seja, os pesos definitivos foram calculados e armazenados para uso nas predições.

8.3 Visualização dos Dados

O módulo de visualização dos dados implementado na CINTIA 2 é simples. O sistema apresenta o gráfico dos espectros para que sejam visualizados e comparados uns com os outros, caso mais de um espectro esteja em classificação ao mesmo tempo. A impressão dos espectros ajuda os pesquisadores mais experientes, que conhecem os padrões dos tipos de SNs, a avaliar se a resposta do classificador condiz com seus conhecimentos prévios. As imagens geradas também são salvas na pasta “Images”.

A execução da visualização de dados é dividida em duas etapas porque a segunda atividade do bloco é a impressão de gráficos de dispersão para que os conjuntos de entradas sejam comparados no espaço de atributos, porém se a quantidade de espectros for pequena a comparação feita com esse tipo de gráfico é irrelevante. Para viabilizar a visualização dos gráficos de dispersão, reduziu-se a dimensionalidade dos espectros para apenas duas dimensões com o método de Isomap, função presente na biblioteca scikit-learn do Python. Isomap é um método não-linear de redução de dimensões que decompõe as entradas dadas em componentes pelo cálculo da distância geodésica entre os pontos vizinhos (TENENBAUM et al., 2000). A Figura 8.3 mostra a visualização dos dados promovida pela CINTIA 2.

Figura 8.3 - Visualização dos dados realizada pela CINTIA 2.



À esquerda a impressão dos espectros e à direita a impressão de um gráfico de dispersão para cada conjunto de entradas, um para cada RNA.

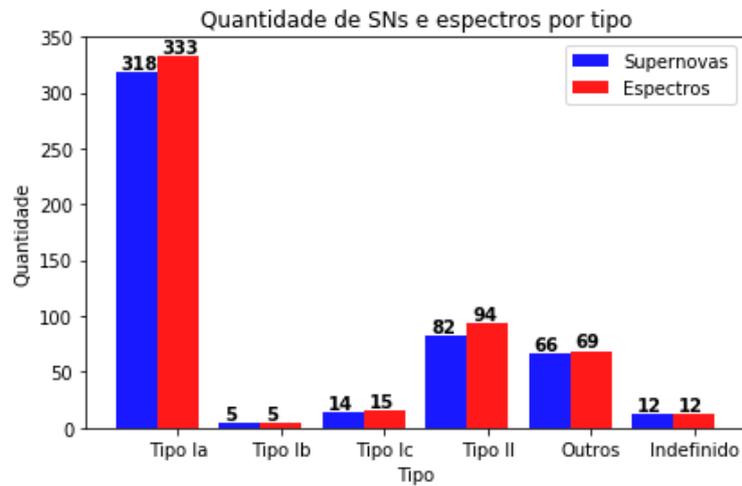
Fonte: Produção da autora.

8.4 Validação do Software

Realizou-se uma etapa de validação da ferramenta que consistiu em submeter novos espectros para a classificação no sistema depois de desenvolvido. Esses espectros são de SNs que não foram usadas nos treinamentos das RNAs. Intentou-se, assim, simular o uso da ferramenta em campo.

Os novos espectros também foram obtidos no TOSC (GILLOCHON et al., 2017), no dia 19 de novembro de 2018. Foram baixados todos os espectros adicionados ao catálogo entre 30 de abril de 2018 e 19 de novembro de 2018. A Figura 8.4 mostra um gráfico que apresenta a quantidade de SNs e espectros baixados por tipo. Dos 528 espectros baixados do catálogo, são considerados apenas 371 para fins de validação, os das categorias “Outros”, “Indefinido” e os espectros de SNs do tipo Ia com fase espectral ≤ -11 e $\geq +11$ dias foram excluídos porque não foram treinados. Depois da submissão ao sistema, 40 espectros foram eliminados no bloco de extração das entradas, pois não possuem o intervalo 4000-7000 Å. Assim, apenas 331 espectros foram efetivamente classificados e podem ser incluídos na contagem da matriz de confusão que resulta nas métricas de avaliação do desempenho.

Figura 8.4 - Quantidade de dados baixados para a etapa de validação da ferramenta.



Além dos 4 tipos que podem ser classificados pela CINTIA 2, há ainda “Outros”, que são tipos peculiares e subtipos que não foram treinados e “Indefinido”, dados sem classificação definida no catálogo.

Fonte: Produção da autora.

A Tabela 8.1 apresenta a matriz de confusão dos quatro módulos da CINTIA 2. Enquanto a Tabela 8.2 mostra as métricas de avaliação de desempenho do classificador. Os valores M são baixos porque os classificadores Ib e Ic não tiveram bom desempenho, dando indícios de que os dois tipos não apresentam padrões consolidados. Porém os valores μ são muito próximos aos dos testes anteriores, evidenciando o bom desempenho da hierarquia de RNAs binárias. O módulo Ia é o que apresenta a melhor performance, corroborando os resultados dos testes dos módulos individuais apresentados no Capítulo 7.

Tabela 8.1 - Matriz de confusão do teste realizado para validar a ferramenta.

	VP	FN	FP	VN
Módulo Ia	196	30	3	102
Módulo Ib	0	5	0	326
Módulo Ic	1	11	1	318
Módulo II	56	32	4	239

Fonte: Produção da autora.

Tabela 8.2 - Métricas de avaliação do desempenho da CINTIA 2 ao ser submetida à validação.

Acurácia Média	0.9350
Taxa de Erro	0.0650
Precisão M	0.6046
Recall M	0.3968
F1-Score M	0.4791
Precisão μ	0.9693
Recall μ	0.7644
F1-Score μ	0.8548

Fonte: Produção da autora.

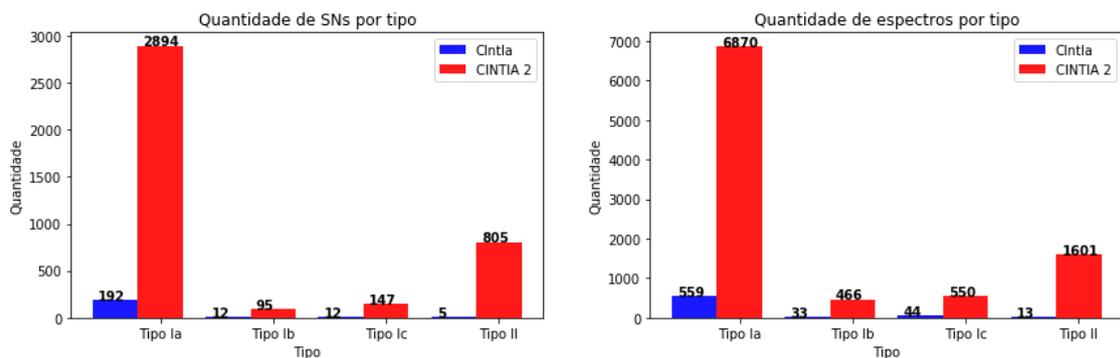
9 DISCUSSÃO SOBRE OS RESULTADOS

Neste capítulo, os resultados alcançados no trabalho são discutidos, comparando a CINTIA 2 com a CIntIa e os outros classificadores descritos no Capítulo 4. Abordar-se as melhorias alcançadas e quais as mudanças de metodologia realizadas para atingi-las.

9.1 CINTIA 2 *versus* CIntIa

A primeira melhoria realizada na CIntIa, transformando-a em CINTIA 2, foi incrementar a quantidade de espectros passíveis de sua análise. A CIntIa trabalhava com espectros de comprimento de onda entre 3800 e 7400 Å, com fase espectral de -3 a +7 dias e filtradas por suavização simples. Essas três características foram alteradas e por causa das mudanças, a quantidade de SNs aumentou mais de 150% (de 221 para 588) e o número de espectros mais de 450% (de 649 para 3697), considerando apenas a base de espectros Cfa, utilizada nos treinamentos e testes da CIntIa. Além das mudanças no cerne da CIntIa, também adicionamos outras bases de dados no processo de treinamento e testes da CINTIA 2. Um gráfico com a comparação da quantidade de dados usada pela CIntIa e pela CINTIA 2 (incluindo todas as bases de dados) é apresentado na Figura 9.1.

Figura 9.1 - Comparativo entre quantidade de SNs (esquerda) e quantidade de espectros (direita) utilizados no desenvolvimento da CIntIa e da CINTIA 2.



Fonte: Produção da autora.

Detalha-se, em seguida, cada uma das mudanças para melhor compreensão das suas

contribuições no aumento da quantidade de dados envolvidos no desenvolvimento da CINTIA 2. Esse aumento é importante por dois motivos principais, o primeiro deles é aumentar a diversidade aprendida pelo sistema, possibilitando seu uso com dados colhidos por diversos instrumentos astronômicos. O outro motivo é tornar a análise que tínhamos anteriormente mais robusta, dado que a quantidade de dados dos tipos não-Ia é pequena para métodos de aprendizagem que necessitam de exemplos, caso dos MLPs.

9.1.1 Mudança no Intervalo de Comprimento de Onda

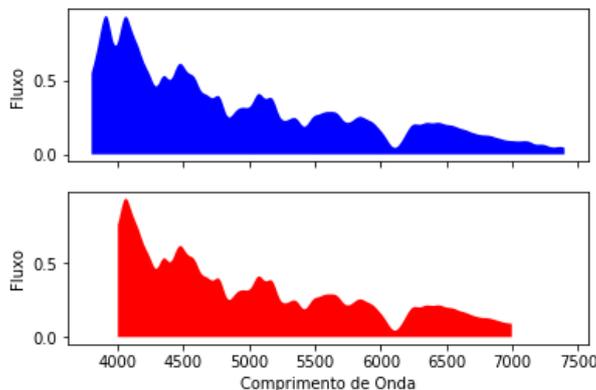
Uma das restrições aplicadas pela CIntIa para classificar os espectros é o intervalo de comprimento de onda em que cada espectro deve estar. Esse intervalo foi definido como 3800-7400 Å, assim a faixa da luz visível está seguramente incluída nas observações. Porém algumas bases de dados contém apenas espectros com amplitudes menores, mas que ainda assim contém a faixa de luz visível. Então, para que não seja necessário aplicar regras diferentes para as diferentes bases de espectros e assim manter a uniformidade no tratamento dos dados, garantindo a automaticidade do sistema, alterou-se o intervalo do comprimento de onda aceito pelo classificador para 4000-7000 Å.

Reduzir o tamanho do intervalo do comprimento de onda provoca o aumento da quantidade de espectros aptos para a classificação, isso acontece porque a exigência feita ao espectro (conter determinado intervalo de comprimento de onda) é mais fácil de ser cumprida. A CINTIA 2 seleciona apenas os espectros que contém o intervalo 4000-7000 Å, faixa do espectro em que está contida as observações de luz visível e onde estão todas as entradas necessárias para as RNAs que efetuam a classificação. A Figura 9.2 exemplifica a mudança, mostrando a seleção feita, em um mesmo espectro, pela CINTIA 2 e pela CIntIa.

9.1.2 Aumento da Amplitude das Fases Espectrais

A fase espectral, que corresponde à idade do espectro contada em dias, é uma informação bastante relevante para a classificação de SNs. Os métodos de classificação mais praticados, automáticos e manuais, baseiam-se em espectros próximos ao período de brilho máximo no dia 0 (fase de máximo). A variação em dias da fase de máximo é pequena, (M6DOLO, 2016) considerou essa fase como -3 a +7 dias, assim todos os espectros, independentemente do tipo, que não estavam nesse intervalo de fase espectral foram descartados da classificação.

Figura 9.2 - Comparação entre os intervalos de comprimento de onda usados, de um espectro da SN1994D, pela CIntIa (acima) e CINTIA 2 (abaixo).



Fonte: Produção da autora.

O alargamento da amplitude das fases espectrais foi possível com a definição de grupos de fases espectrais próximas que formam períodos em que os espectros apresentem características similares (Tabela 7.2). Assim o módulo da CINTIA 2 responsável por identificar os espectros de SNs Ia é capaz de reconhecer espectros em fases além do que era possível com a CIntIa, o limite inferior foi de -3 para -10.9 e o limite superior de +7 para +10.9. Enquanto todos os outros módulos foram treinados para reconhecer sem restrições de fases espectrais.

Na Tabela 9.1, é feita uma comparação entre os números da CIntIa, da CINTIA 2 e o treinamento 1 da Tabela 7.3, onde foram utilizados espectros de SNs Ia obtidos no TOSC apenas na fase de máximo (> -4 e $< +4$). Os treinamentos usados para comparar são provenientes da RNA que classifica em tipo Ia ou não-Ia. Observa-se que os resultados das métricas de avaliação baixam quando a quantidade de dados e as fases espectrais aceitas são expandidas. No entanto, houve a preocupação em manter a Acurácia ≥ 0.90 e o Índice Kappa > 0.80 , assim o classificador foi mantido com concordância de classificação “Quase Perfeita”, enquanto a quantidade de espectros do tipo Ia aumentou em mais de 500% e as do tipo não-Ia em mais de 2600%.

Tabela 9.1 - Comparação entre quantidade de espectros, fases espectrais e principais métricas de avaliação da RNA Ia da CIntIa; o treino chamado de Intermediário Ia, cujos espectros do tipo Ia estão na fases > -4 e $< +4$ dias; e o módulo Ia da CINTIA 2.

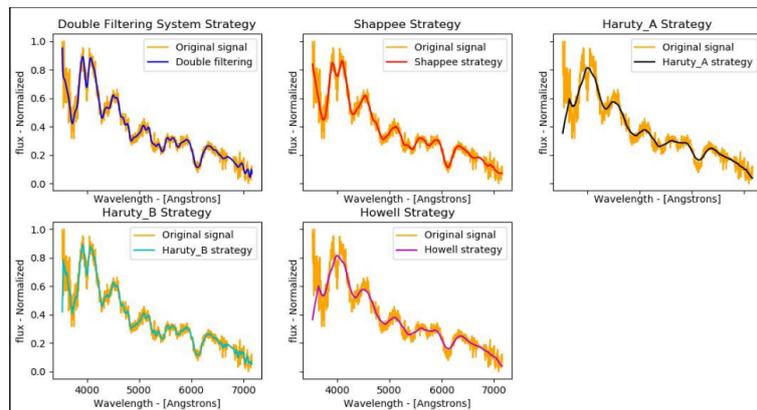
	CIntIa	Intermediário Ia	CINTIA 2
Espectros Ia	559	1418	3455
Espectros não-Ia	90	2459	2459
Fases Ia (em dias)	-3 a +7	-3.9 a +3.9	-10.9 a +10.9
Acurácia	0.9915	0.9530	0.9196
Precisão	0.9907	0.9813	0.9951
Kappa	0.9518	0.8978	0.8385

Fonte: Produção da autora.

9.1.3 Mudança no Método de Filtragem dos Espectros

Uma das etapas mais importantes do pré-processamento é a filtragem do sinal original obtido pelo espectroscópio. O método de filtragem empregado na CIntIa é a suavização por Médias Móveis, que calcula uma série de médias do conjunto total e seus subconjuntos para eliminar os ruídos existentes. Porém, (ARANTES FILHO; GUIMARÃES, 2017) argumentam que outros métodos podem ser mais vantajosos na filtragem de espectros de luz provenientes de SNs.

Figura 9.3 - Comparação entre estratégias de filtragem de espectros de SNs, aplicadas em um espectro da SN1998dx.



Fonte: Arantes Filho e Guimarães (2017).

(ARANTES FILHO; GUIMARÃES, 2017) apresentam uma comparação entre os filtros utilizados nos trabalhos consultados, que tratam especificamente de espectros de SNs. Além disso, os autores também propõem a Dupla-Filtragem com o filtro Savitzky-Golay, utilizada neste trabalho. A Dupla-Filtragem se mostra superior em diversas métricas, inclusive na correlação entre os sinais filtrados pelo método e os originais. A Figura 9.3 é uma comparação visual entre a filtragem feita pela Dupla-Filtragem e outras cinco estratégias no espectro da SN1998dx. A estratégia chamada de “Haruty_A” é a mesma utilizada na CIntIa.

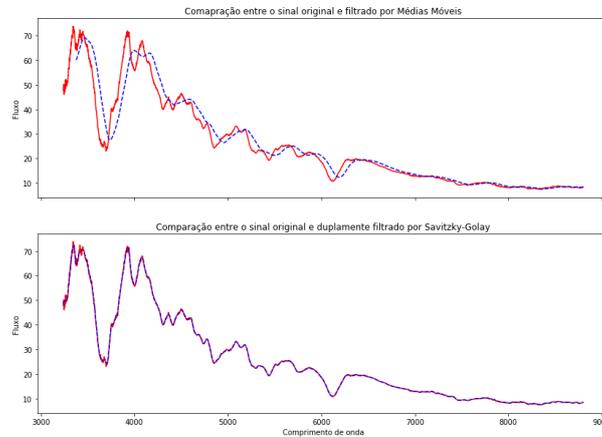
Visualmente, percebe-se que a Dupla-Filtragem preserva melhor o sinal, eliminando as regiões muito ruidosas, do que a Haruty_A. As métricas apresentadas na Tabela 9.2, confirmam a experiência visual. As métricas escolhidas para avaliação das estratégias são medidas de correlação (correlação cruzada, Pearson e Spearman), que correlacionam o sinal original de cada espectro do conjunto de testes com o seu correspondente sinal filtrado; e medidas que medem a distância de um sinal para outro (distância entre os sinais e raiz quadrada do erro médio). A superioridade da Dupla-Filtragem em todas as medidas asseguram que a filtragem feita com esta estratégia resulta em espectros mais adequados para o uso posterior. Logo, espectros que não foram usados no desenvolvimento da CIntIa por serem ruidosos em demasia, cuja recuperação não foi possível com a técnica de filtragem por Médias Móveis, não foram descartados pela CINTIA 2 que utiliza a Dupla-Filtragem com o filtro Savitzky-Golay. A Figura 9.4 fornece um outro exemplo da filtragem do mesmo espectro (da SN1994D) com a técnica de Médias Móveis utilizada pela CIntIa e a Dupla-Filtragem que faz parte do pré-processamento da CINTIA 2.

Tabela 9.2 - Comparação entre métricas de avaliação dos métodos de filtragem Dupla-Filtragem e Médias Móveis aplicados a espectros de SNs.

	Dupla-Filtragem	Haruty_A (Médias Móveis)
Correlação cruzada	99.63%	94.85%
Pearson	98.47%	78.68%
Spearman	98.81%	81.20%
Distância entre os sinais	0.0153	0.2130
Raiz Quadrada do Erro Médio	0.0369	0.1400

Fonte: Adaptado de Arantes Filho e Guimarães (2017).

Figura 9.4 - Comparação entre a filtragem de um espectro da SN1994D feita por Médias Móveis (acima) e Dupla-Filtragem Savitzky-Golay (abaixo).



O sinal original está na cor vermelha e os sinais de filtragem na cor azul.

Fonte: Produção da autora.

9.2 CINTIA 2 *versus* Outros Classificadores

Uma comparação entre a CINTIA 2 e os outros classificadores consultados na literatura é importante para avaliar em que aspectos ela representa uma melhoria e depreender o quão vantajosa é a sua utilização. A Tabela 9.4 apresenta os critérios escolhidos para efetuar essa verificação (a Tabela 9.3 apenas auxilia na em sua formatação). Considera-se se o classificador usa Inteligência Computacional ou não, a quantidade de espectros utilizados, a amplitude das fases espectrais tratadas e as métricas de avaliação Acurácia, Precisão, Recall e *F1-Score*. A sigla “NA” significa não aplica, ou seja, a informação não integra o escopo do trabalho. Enquanto “NC” é não consta, quando a informação não aparece no trabalho. Ressalta-se que os testes realizados, expressos nas métricas de avaliação, não usam o mesmo conjunto de espectros. Cada classificador foi testado com um conjunto de dados melhor detalhado em sua respectiva referência.

Tabela 9.3 - Identificadores dos classificadores automáticos para utilização na Tabela 9.4.

Identificador	Classificador
1	SNID
2	GELATO
3	DRACULA
4	Quantitative
5	SUZAN

Fonte: Produção da autora.

Tabela 9.4 - Comparação entre CINTIA 2 e outros classificadores automáticos de SNs. "NA"significa não aplica e "NC", não consta. "IA"significa Inteligência Artificial.

	1	2	2	4	5	CINTIA 2
IA?	Não	Não	Sim	Não	Sim	Sim
Espectros Ia	879	1009	486	146	3082	6870
Espectros Ib	NA	129	NA	12	217	466
Espectros Ib/c	322	83	NA	4	NA	NA
Espectros Ic	NA	355	NA	19	282	550
Espectros II	353	1132	NA	NA	116	1601
Menor fase	-15	-15	-3	-5	-15	-1435
Maior fase	+70	+600	+3	+5	+2959	+8303
Acurácia	NC	NC	NC	0.95	0.73	0.95
Precisão	NC	NC	NC	NC	0.93	0.97
Recall	NC	NC	NC	NC	0.72	0.76
F1-Score	NC	NC	NC	NC	0.83	0.85

Fonte: Produção da autora.

Ao analisar-se a Tabela 9.4, no que concerne à quantidade de dados e a abrangência das fases espectrais, observa-se que a CINTIA 2 é superior a todos os outros classificadores, ou seja, ela compreende uma diversidade maior de espectros. A CINTIA 2 transcende todos os classificadores consultados, que apresentam bons resultados apenas para espectros na fase de brilho máximo. Considerando que a expectativa é tornar o sistema um classificador atuante em conjunto com telescópios, essa é uma característica importante. A sua Acurácia Média e a Precisão μ (apresentadas na Tabela 7.18) são equiparáveis aos sistemas que também apresentam essa informação. Como as métricas de avaliação não são fornecidas por todos os autores dos classifi-

cadores, não é possível fornecer uma comparação mais profunda de desempenho. No entanto, é possível perceber que a CINTIA 2 consegue processar mais dados pois seu treinamento incluiu uma quantidade de espectros muito superior a todos os outros classificadores.

9.3 Considerações sobre a Estratégia Um Contra Todos

A estratégia Um Contra Todos é utilizada para decompor um problema de multi-classe complexo em problemas menores, cada um dos subproblemas fica com duas classes. Apesar da CINTIA 2 lidar com apenas quatro classes, (MóDOLO, 2016) apresenta resultados que mostram que decompor esse problema, de classificar espectros de SNs em quatro classes, é melhor do que utilizar apenas uma RNA. Em seu experimento, feito durante o desenvolvimento da CIntIa, uma RNA com 376 entradas treinou 265 padrões e testou 66, obtendo no máximo Acurácia de 75.80%. Os resultados expressivos da CIntIa na classificação de espectros de SNs Ia foram alcançados após a aplicação da fase de decomposição que originou quatro RNAs binárias.

A fase de reconstrução, que não foi implementada na CIntIa e realizada neste trabalho, não usou a estratégia clássica de votação distribuída. Optou-se por definir a reconstrução por meio de uma hierarquia de RNAs, onde a RNA com melhor resultado individual classifica o espectro e apenas se o espectro não for reconhecido por ela a classificação da RNA seguinte é solicitada, se for reconhecido o espectro é imediatamente classificado.

Experimentou-se fazer a reconstrução por votação distribuída para fins de comparação. Uma das desvantagens em relação a hierarquia implementada é que algumas classes podem empatar na classificação de um espectro e isso requer uma outra estratégia combinada para resolver o impasse. A Tabela 9.5 apresenta as matrizes de confusão desses experimentos, acrescidas de uma coluna com a quantidade de empates ocorridos com os dados da validação da ferramenta, e os valores obtidos com o emprego da hierarquia, que são os mesmos da Tabela 8.1. Percebe-se que a quantidade de acertos das RNAs Ia e II é mais ampla quando aplicada a hierarquia de RNAs na reconstrução da decomposição aplicada no problema de multi-classe tratado. Além disso, quando a hierarquia é usada não há casos de empate, todas as classificações são efetuadas de forma única, por outro lado ao utilizar a votação distribuída algumas classificações terminaram em empate porque duas RNAs reconheceram o padrão como sendo de sua classe positiva. Portanto, entende-se que a hierarquia de RNAs usada para definir a classificação final do espectro, após ele passar pelo reconhecimento das quatro RNAs binárias, é mais adequada do que a

votação distribuída, método bastante usado e descrito na literatura.

Tabela 9.5 - Comparação entre as matrizes de confusão de um experimento que aplica votação distribuída na fase de reconstrução da estratégia Um Contra Todos e outro que aplica a hierarquia de RNAs definida neste trabalho.

	VP	FN	FP	VN	Empates
Hierarquia Ia	196	30	3	102	0
Hierarquia Ib	0	5	0	326	0
Hierarquia Ic	1	11	1	318	0
Hierarquia II	56	32	4	239	0
Votação Distribuída Ia	194	31	1	104	1
Votação Distribuída Ib	0	5	0	326	0
Votação Distribuída Ic	1	11	1	318	0
Votação Distribuída II	55	30	4	239	3

Fonte: Produção da autora.

10 CONSIDERAÇÕES SOBRE A CLASSIFICAÇÃO DE FASES ESPECTRAIS DE SUPERNOVAS Ia

A classificação de fases espectrais de SNs Ia consiste na identificação da idade dos espectros de uma SNIa. Neste trabalho é abordada a classificação automática promovida por (RIESS et al., 1997), que introduz um método matemático para descobrir a idade de espectros de SNs Ia usando apenas o próprio espectro como informação, independentemente da curva de luz. “A evolução temporal do espectro de supernovas Ia provê uma forma alternativa e mais confiável de medir a passagem do tempo por supernovas Ia individuais”, explicam os autores. Os espectros são divididos em 8 regiões do comprimento de onda chamados de *Spectral Feature Age* (SFA), cada região corresponde a uma ou mais linhas espectrais que se manifestam em espectros de SNs Ia, como mostrado na Tabela 10.1. As *features* mais adequadas para calcular a fase são as de 2 a 5 porque são regiões cujos elementos sofrem mudanças mais rapidamente, característica das regiões que contêm Fe II.

Tabela 10.1 - Intervalos do comprimento de onda usadas como SFA para calcular fases espectrais.

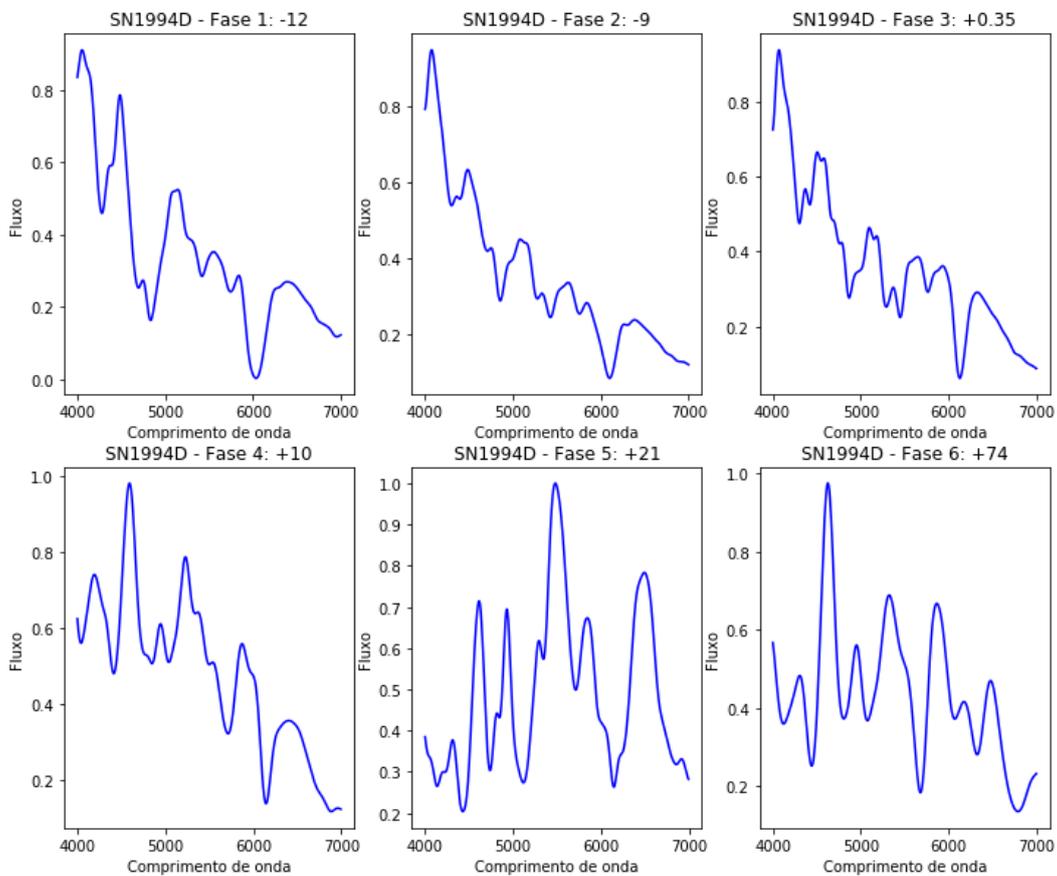
Feature	Intervalo (em Å)	Elementos
1	3800-4200	Si II, Ca II
2	4200-4580	Mg II, Fe II
3	4580-4950	Fe II
4	4950-5200	Fe II
5	5200-5600	S II
6	5600-5900	Si II, Na II
7	5900-6300	Si II
8	6300-6800	Fe II

Fonte: Adaptado de Riess et al. (1997).

O método realizado por (RIESS et al., 1997), bem como o implementado por (BLONDIN; TONRY, 2007), para identificar a fase espectral de um espectro individualmente faz isso calculando uma idade absoluta (numérica). A partir do trabalho de (ARANTES FILHO, 2018), que sugere a divisão das fases em quatro intervalos, expandiu-se essa divisão para seis intervalos (Tabela 7.2), chamados de blocos de fases espectrais. A divisão em blocos foi feita para que o método de Inteligência Artificial, como as RNAs cujas saídas são neurônios que podem apresentar apenas saídas binárias, seja aplicado adequadamente. Na Figura 10.1, mostra-se seis espectros da SN1994D, con-

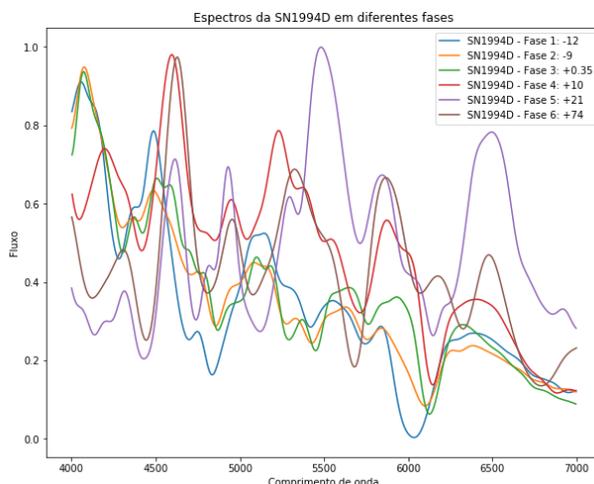
siderada uma SNIa padrão que segundo (PATAT et al., 1995) foi classificada como do tipo Ia antes de chegar ao brilho máximo. Cada um dos espectros está em um bloco de fases espectrais definido na Tabela 7.2 e a fase absoluta está no título de cada sub-figura. A Figura 10.2 dispõe os seis espectros em uma mesma região (4000-7000 Å) para que a comparação visual seja mais explícita, assim é possível observar as diferenças que ocorrem com a passagem do tempo.

Figura 10.1 - Espectros da SN1994D em 6 fases espectrais diferentes. Cada fase está em um bloco de fases de acordo com a Tabela 7.2.



Fonte: Produção da autora.

Figura 10.2 - Espectros da SN1994D mostrados no mesmo espaço para comparação das mudanças nas linhas espectrais no decorrer do tempo.



Fonte: Produção da autora.

Respeitando os intervalos de comprimento de onda descritos por (RIESS et al., 1997) como os que modificam mais rapidamente, intentou-se a classificação das fases com métodos de Inteligência Artificial. O primeiro deles foram as RNAs binárias, mesma estratégia discutida em todo o trabalho para classificação dos tipos e o segundo foi o método K-Means, que classifica de forma não-supervisionada. As *features* escolhidas foram as 2, 3 e 4, a *feature* 5 não foi usada porque o intervalo correspondente faz parte do intervalo usado na classificação de tipos, a qual busca semelhanças entre os padrões de SNs Ia, na classificação de fases interessa as diferenças. As seções seguintes exploram os dois experimentos realizados e seus resultados.

10.1 Classificação com RNAs Binárias

Dividiu-se os espectros de SNs Ia em 80% para treinamento e 20% para teste. A Tabela 10.2 mostra a quantidade de espectros destinadas para cada etapa de classificação, separados por fase espectral. Todos os espectros foram previamente pré-processados, como descrito na Seção 5.2. O intervalo usado foi de 4200 a 5200 Å, que corresponde as *features* 2, 3 e 4 (Tabela 10.1). Após o pré-processamento, os espectros passaram por uma interpolação a cada 8 pontos para igualar a quantidade de pontos de todos os espectros, resultando em 126 valores de fluxo utilizados como entradas. Os parâmetros de treinamento das seis RNAs binárias, uma para cada

classe (bloco de fases), são os mesmos da Seção 7.2. E as topologias testadas estão na Tabela 7.1.

Tabela 10.2 - Quantidade de espectros utilizados na classificação das fases espectrais de SNs Ia.

	Espectros treinamento	Espectros teste
Pré-Pré-Máximo	230	58
Pré-Máximo	556	139
Máximo	1185	296
Pós-Máximo	1137	275
Pré-Nebular	1658	414
Nebular	557	139

Fonte: Produção da autora.

A Tabela 10.3 exibe as matrizes de confusão dos melhores resultados de cada uma das seis RNAs treinadas. E a Tabela 10.4 apresenta as métricas de avaliação do desempenho. Pode-se observar que essa abordagem não é satisfatória para todo o conjunto, apenas as fases mais avançadas (a partir de +11 dias) tem resultados que evidenciam alguma capacidade de se distinguir das outras fases. Apesar da mudança dos valores das entradas, as RNAs que classificam as fases Pré-Máximo, Máximo e Pós-Máximo tem os menores Índices Kappa, corroborando os resultados da classificação por tipos de espectros de SNs Ia obtido neste trabalho. Ou seja, a diferenciação dos espectros das três fases não é simples, por isso foi possível ampliar o intervalo de fases espectrais que o Módulo Ia da CINTIA 2 é capaz de reconhecer.

Tabela 10.3 - Matrizes de confusão dos testes das seis RNAs binárias que classificam as fases de espectros de SNs Ia.

	VP	FN	FP	VN
Pré-Pré-Máximo	2	56	0	1263
Pré-Máximo	2	137	2	1180
Máximo	3	193	2	1023
Pós-Máximo	1	274	0	1046
Pré-Nebular	83	331	15	892
Nebular	124	15	260	922

Fonte: Produção da autora.

Tabela 10.4 - Métricas de avaliação do desempenho das seis RNAs binárias que classificam as fases espectrais de SNs Ia. Os mnemônicos que designam as RNAs tem as seguintes significações. PrePreMax: Pré-Pré-Máximo; PreMax: Pré-Máximo; Max: Máximo; PosMax: Pós-Máximo; PreNeb: Pré-Nebular; Neb: Nebular.

	PrePreMax	PreMax	Max	PosMax	PreNeb	Neb
Acurácia	0.9576	0.8948	0.8403	0.7926	0.7381	0.7918
Precisão	1	0.5000	0.6000	1	0.8469	0.3229
Recall	0.0345	0.0144	0.0153	0.0036	0.2005	0.8921
F1-Score	0.0667	0.0280	0.0298	0.0072	0.3242	0.4742
Especificidade	1	0.9983	0.9980	1	0.9835	0.7800
AUC	0.5172	0.5063	0.5067	0.5018	0.5920	0.8361
Kappa	0.0639	0.0222	0.0220	0.0057	0.2321	0.3781

Fonte: Produção da autora.

10.2 Classificação com K-Means

O K-Means é um algoritmo de agrupamento não-supervisionado. Como algoritmo de agrupamento, entende-se um método para formação de subconjuntos dentro de um conjunto de dados maior nos quais os dados agrupados partilham semelhanças. O agrupamento é não-supervisionado porque não é necessário fornecer exemplos das saídas esperadas, o algoritmo forma os agrupamentos sem auxílio externo. Como explicam (JAMES et al., 2013), “para executar o agrupamento por meio do K-Means, é necessário primeiramente especificar o número K de grupos ; então o algoritmo K-Means atribuirá cada observação a exatamente um dos K grupos.” O algoritmo K-Means é apresentado no Algoritmo 1.

início

Atribuir aleatoriamente um número, de 1 a K, para cada um dos padrões de entrada. Eles servem como atribuições iniciais de agrupamento para as observações;

repita

Para cada um dos grupos K, calcule o centróide do grupo;
 Atribuir cada observação ao grupo cujo centróide está mais próximo (onde o mais próximo é definido usando a distância euclidiana);

até a atribuição de centróides parar de mudar;

fim

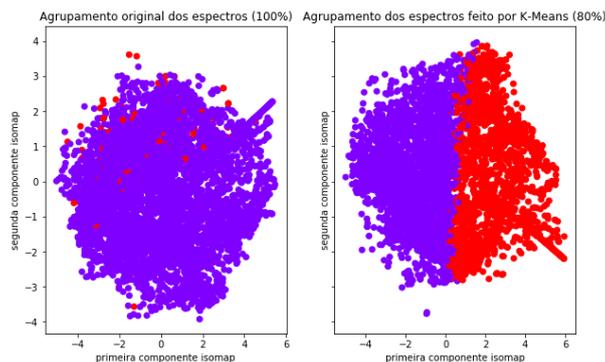
Fonte: Adaptado de James et al. (2013).

Algoritmo 1: Agrupamento com K-Means

Assim como na classificação com RNAs binárias, todos os espectros foram previamente pré-processados, como descrito na Seção 5.2. O intervalo usado foi de 4200 a 5200 Å, que corresponde as *features* 2, 3 e 4 (Tabela 10.1). Após o pré-processamento, os espectros passaram por uma interpolação a cada 8 pontos para igualar a quantidade de pontos de todos os espectros, resultando em 126 valores de fluxo utilizados como entradas. Os dados foram divididos em 80% para treinamento e 20% para teste. Usou-se o algoritmo K-Means implementado pela biblioteca scikit-learn, desenvolvida para Python, cujo único parâmetro obrigatório é a quantidade de grupos que se deseja formar. Executou-se o algoritmo 6 vezes com quantidade de grupos igual a 2, cada execução do algoritmo faz uma classificação binária, assim como as RNAs da seção acima.

As Figuras de 10.3 a 10.8 apresentam a comparação entre a distribuição espacial, em gráficos de dispersão, da classificação binária real dos espectros, realizada por especialistas, e a classificação realizada pelo algoritmo K-Means. Reduziu-se a dimensão para duas com Isomap apenas para fins de visualização dos dados. Observa-se que os dados em todos os casos estão total ou parcialmente sobrepostos em escala 2D, resultando, assim na dificuldade de separação por métodos como o K-Means e até mesmo como o Perceptron de Múltiplas Camadas, que foi experimentado anteriormente.

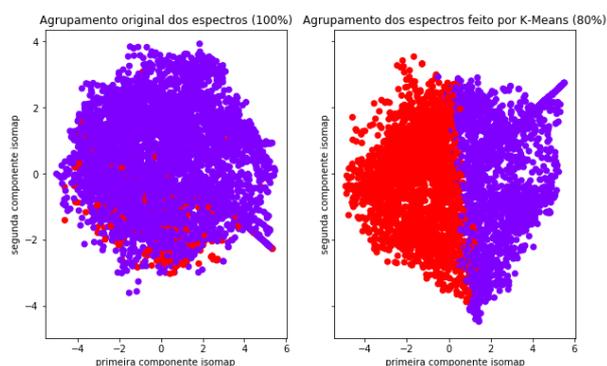
Figura 10.3 - Gráfico de dispersão de espectros de SNs na fase de pré-pré-máximo e não-pré-pré-máximo.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

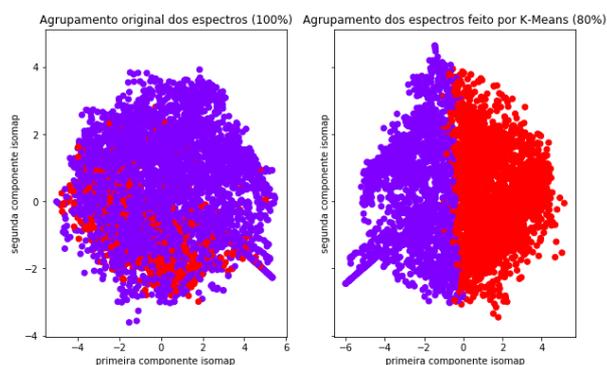
Figura 10.4 - Gráfico de dispersão de espectros de SNs na fase de pré-máximo e não-pré-máximo.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

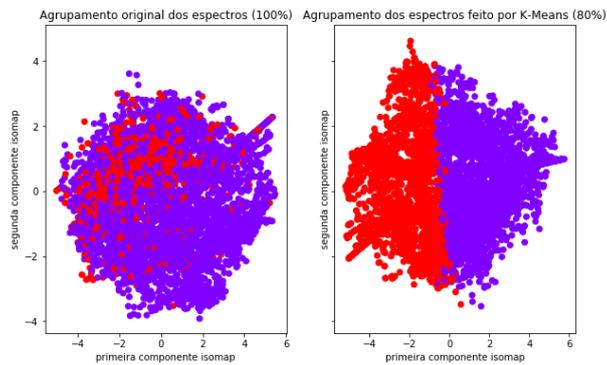
Figura 10.5 - Gráfico de dispersão de espectros de SNs na fase de máximo e não-máximo.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

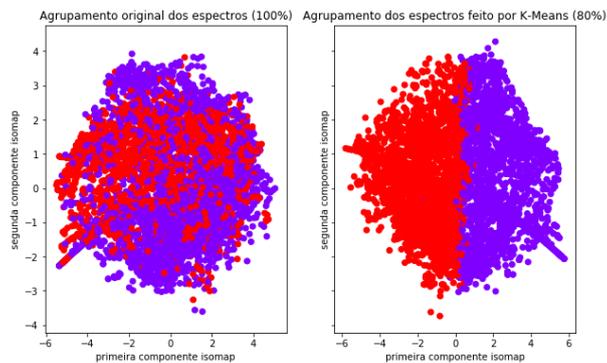
Figura 10.6 - Gráfico de dispersão de espectros de SNs na fase de pós-máximo e não-pós-máximo.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

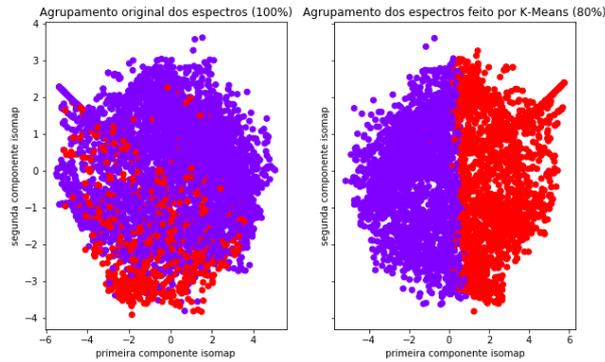
Figura 10.7 - Gráfico de dispersão de espectros de SNs na fase de pré-nebular e não-pré-nebular.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

Figura 10.8 - Gráfico de dispersão de espectros de SNs na fase de nebular e não-nebular.



À esquerda é mostrada a divisão em dois grupos com a classificação real de todo o conjunto e à direita, a divisão em dois grupos feito pelo algoritmo K-Means de 80% do conjunto (dados de treinamento).

Fonte: Produção da autora.

As Tabelas 10.5 e 10.6 exibem as matrizes de confusão de cada classificação binária realizada com K-Means e as métricas de avaliação do desempenho. Os resultados expressos nas tabelas confirmam a inadequação do método K-Means para a classificação proposta, como visualizado nos gráficos. Nenhum dos classificadores apresenta bons resultados. Em geral, as melhores métricas são as do classificador que separa entre os espectros na fase pós-máximo e fora da fase pós-máximo, ainda assim não se destaca muito dos outros resultados.

Tabela 10.5 - Matrizes de confusão das classificações binárias realizadas pelo algoritmo K-Means.

	VP	FN	FP	VN
Pré-Pré-Máximo	40	18	515	756
Pré-Máximo	40	77	724	488
Máximo	102	231	600	396
Pós-Máximo	222	76	501	530
Pré-Nebular	239	178	578	334
Nebular	60	71	449	749

Fonte: Produção da autora.

Tabela 10.6 - Métricas de avaliação do desempenho dos seis classificadores binários implementados com K-Means.

	PrePreMax	PreMax	Max	PosMax	PreNeb	Neb
Acurácia	0.5889	0.3973	0.3747	0.5658	0.4312	0.6087
Precisão	0.0721	0.0524	0.1453	0.3071	0.2925	0.1179
Recall	0.6879	0.3419	0.3063	0.7450	0.5731	0.4580
F1-Score	0.1305	0.0908	0.1971	0.4349	0.3874	0.1875
Especificidade	0.5948	0.4026	0.3976	0.5141	0.3662	0.6252
AUC	0.6422	0.3723	0.3519	0.6295	0.4997	0.5416
Kappa	0.0559	-0.0730	-0.2163	0.1719	-0.0481	0.0364

Fonte: Produção da autora.

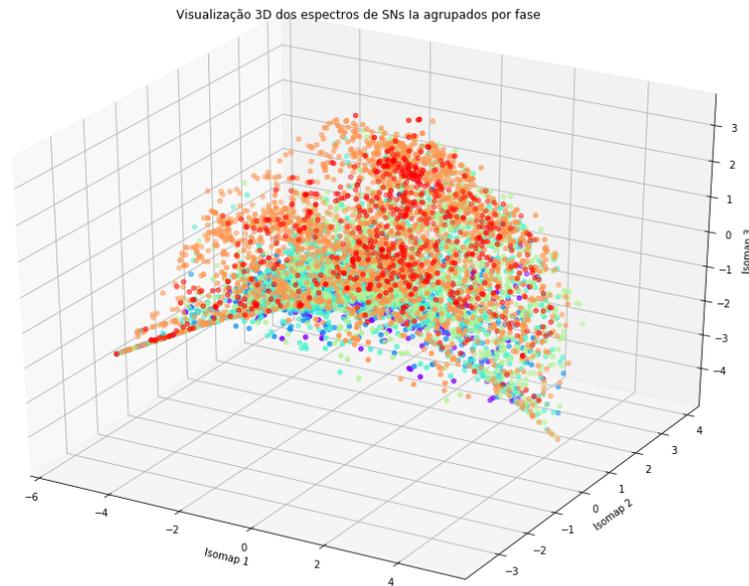
10.3 Dificuldades de Classificação das Fases Espectrais

Especialistas em SNs afirmam que a mudança dos espectros no decorrer do tempo ocorre, logo é esperado que seja possível classificá-los de acordo com as fases espectrais. Usando RNAs do tipo Perceptron é possível separar os padrões por planos ou hiper-planos (OSÓRIO, 1999), porém como demonstrado empiricamente, a classificação das fases espectrais de SNs do tipo Ia é um problema de difícil solução por esse método. O hiperplano formado para separar os dados, cuja equação não é conhecida, pois o aprendizado de uma RNA funciona como uma caixa preta, não se mostra satisfatório para as condições apresentadas. Os resultados insatisfatórios do método podem acontecer devido a algumas razões: conjunto de entradas inadequado; necessidade de mais ou menos camadas e neurônios, que não foram testados; taxa de aprendizagem ou taxa de *momentum* impróprias. Novos experimentos podem ser realizados no futuro a fim de observar o desempenho do método com novos parâmetros.

O método K-Means forma agrupamentos de dados de acordo com o cálculo da distância euclidiana a determinadas centróides, o que se observa nos gráficos de dispersão (Figuras 10.3 a 10.8) é que os dados estão todos muito próximos o que dificulta a separação proporcionada pelo algoritmo, feita por retas. A inadequação das entradas pode ser uma das causas, bem como a forma de inicializar as centróides utilizadas pelo algoritmo. Um outro trabalho que pode ser feito futuramente é a mudança do conjunto de entradas, como outras medidas que descrevam o espectro, e a variação da quantidade de centróides, podendo inclusive resultar em outros intervalos para definição das fases. Para ilustrar a baixa dispersão dos espectros, considerando os

126 valores de fluxo do intervalo 4200-5200 Å interpolados a cada 8 pontos, que foram usados como entradas, a Figura 10.9 apresenta um gráfico de dispersão em 3D usando as três componentes principais obtidas com o algoritmo Isomap.

Figura 10.9 - Gráfico de dispersão 3D dos espectros separados em seis fases espectrais, cada dimensão é um componente principal obtida por meio do algoritmo Isomap.



Fonte: Produção da autora.

11 CONCLUSÃO

Neste trabalho, foi apresentada uma nova versão para o Classificador Inteligente de Supernovas tipo Ia (CIntIa), proposto por (MÉDOLO, 2016). A nova versão é chamada de CINTIA 2, mostra-se os aspectos de concepção e implementação das melhorias nesse sistema. A CINTIA 2 é um sistema baseado em uma hierarquia de Redes Neurais Artificiais do tipo Perceptron de Múltiplas Camadas, que classifica os espectros de supernovas dos tipos Ia, Ib, Ic e II, com ênfase nas do tipo Ia. A hierarquia é composta por quatro redes neurais (módulos) que fazem classificações binárias, cada módulo identifica um tipo de supernova.

Foram experimentadas sete topologias diferentes para cada rede neural e cada uma delas passou por dois testes com conjuntos de dados distintos. A classificação dos tipo Ia e II obtiveram resultados excelentes na identificação dos valores positivos e negativos, enquanto os tipos Ib e Ic não apresentam resultados tão animadores. A CINTIA 2 foi implementada nas linguagens de programação C++ e Python e é um software pronto para uso em conjunto com instrumentos que coletam os espectros de luz das supernovas. O sistema automatiza desde o pré-processamento dos espectros até a classificação propriamente dita.

Algumas contribuições que este trabalho traz podem ser destacadas. A primeira delas é o aumento da quantidade de dados usados nas fases de treinamento e testes. Incrementou-se a diversidade de dados aprendida pelas redes neurais com a adição dos espectros obtidos no The Open Supernova Catalog (GILLOCHON et al., 2017) que reúne os acervos de 17 bancos de espectros gratuitos e contribuições individuais de pesquisadores. A maior quantidade de dados, em comparação aos que foram analisados pela CIntIa, permitiu que a análise realizada fosse mais robusta. A generalização feita pelo classificador agora engloba dados provenientes de uma grande variedade de telescópios e espectroscópios com características diferentes, características essas que podem acarretar mudanças entre os espectros.

Outras duas contribuições importantes dizem respeito ao pré-processamento dos espectros. A primeira delas foi a adoção da Dupla-Filtragem, proposta por (ARANTES FILHO; GUIMARÃES, 2017). A Dupla-Filtragem é a aplicação do filtro de Savitzky-Golay duas vezes seguidas em cada espectro, essa estratégia se mostrou mais eficaz na eliminação de ruídos do que a suavização por Médias Móveis, aplicada anteriormente pela CIntIa. Diminuiu-se o intervalo de comprimento de onda avaliado pelo sistema de 3800-7400 Å para 4000-7000 Å, assim aumentou-se a quantidade de dados passível de classificação sem comprometer o intervalo usado como entrada para

as RNAs, que corresponde à faixa de luz visível.

A CIntIa fazia classificações de espectros que estivessem apenas em torno do brilho máximo da SN, definido como -3 a +7 dias. Logrou-se expandir essa amplitude de fases espectrais, a CINTIA 2 é capaz de identificar espectros de SNs Ia entre -10.9 e +10.9 dias e para os outros tipos foram incluídos nos treinamentos espectros em qualquer fase espectral. Essa contribuição é bastante importante porque o desenvolvimento da CINTIA 2 visa fornecer um classificador que possa ser usado de forma automática e em tempo real, conjuntamente com instrumentos localizados em regiões remotas, como os do projeto chinês KDUST. Assim, a classificação da SN pode ser feita imediatamente, sem a necessidade de esperar a curva de luz ser traçada para que os espectros adequados sejam escolhidos e submetidos ao classificador.

Propôs-se, também, uma arquitetura que integra as quatro redes neurais treinadas em um só sistema, dessa forma, eliminamos a possibilidade de classificações ambíguas. A hierarquia foi realizada com o intuito de prover uma fase de reconstrução para o método Um Contra Todos utilizado anteriormente apenas na fase de decomposição. A ordem das RNAs na hierarquia foi escolhida de acordo com o desempenho de cada uma delas nos testes a que foram submetidas para verificar a qualidade do treinamento. Portanto, primeiro é acionado o módulo que identifica o tipo Ia, depois o módulo II, em seguida os módulos Ib e Ic, respectivamente. Caso o espectro não seja classificado em nenhum dos tipos ele recebe o rótulo de “Tipo Não-Identificado” e assim pode receber outro tipo de atenção dos especialistas que monitoram o funcionamento da base de observação remota.

Como citado anteriormente, a hierarquia foi implementada em um sistema que atualmente se encontra pronto para uso. A CINTIA 2 realiza o pré-processamento do espectro ainda em estado bruto, necessitando saber apenas o valor de *redshift* associado, mostra os dados em um módulo simples de visualização de dados e efetua a classificação. O módulo de visualização de dados exibe e salva os gráficos dos espectros e gráficos de dispersão, caso mais de 10 espectros estejam sendo classificados ao mesmo tempo, que permitem a comparação visual dos espectros. O resultado da classificação também é exibido na tela e salvo, o que favorece o envio de resultados para outros pesquisadores caso o operador entenda ser necessário. O software foi validado com espectros de SNs não utilizados nos treinamentos, o que confirmou o bom desempenho dos módulos Ia e II e a dificuldade em classificar os do tipo Ib e Ic.

Incluiu-se neste trabalho ainda, um capítulo que trata da classificação de espectros

de SNs Ia em fases espectrais. Foi proposta uma divisão das fases em seis grupos e mostrados os resultados dos experimentos de classificação por RNAs binárias e pelo algoritmo K-Means, também usando a estratégia de binarização dos classificadores. Os resultados obtidos não foram satisfatórios, mas foram destacadas algumas razões que podem ser as causadoras do insucesso e sugeridos trabalhos a serem efetuados futuramente para a realização dessa tarefa.

Entende-se que alguns trabalhos podem ser derivados deste. Futuramente, a implementação de uma estratégia de escolha automática de configurações para as RNAs é uma evolução importante, dessa forma a quantidade de treinamentos é expandida possibilitando melhorias nos resultados. O algoritmo de treinamento das RNAs também necessita de melhorias na estratégia de parada, como o uso de validação cruzada para encontrar pontos de parada que resultem em treinamentos mais acurados. No que concerne à classificação das fases espectrais, deve-se investigar outros conjuntos de entrada para os métodos de aprendizagem utilizados. Experimentar outros métodos de aprendizagem de máquina, além de RNAs Perceptron e K-Means, também é um estudo que deve ser considerado, já que a quantidade de trabalhos com essa temática não é vasto. Além disso, pode ser interessante estabelecer treinamentos para redes neurais específicas que reconheçam os subtipos do tipo Ia (tipos peculiares) o que não foi abordado neste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ABAZAJIAN, K. et al. The first data release of the sloan digital sky survey. **The Astronomical Journal**, v. 126, p. 2081–2086, out. 2003. 29

ALDERING, G. et al. Overview of the nearby Supernova factory. In: TYSON, J. A.; WOLFF, S. (Ed.). **Survey and other telescope technologies and discoveries**. [S.l.: s.n.], 2002. (), p. 61–72. 29

ARANTES FILHO, L. R. **Classificação inteligente de supernovas utilizando sistemas de regras nebulosas**. Dissertação (Mestrado em Computação Aplicada) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2018. 1, 26, 27, 28, 32, 33, 35, 75

ARANTES FILHO, L. R.; GUIMARÃES, L. N. F. Double filtering system for analysis and processing of supernovae spectral data. In: SIMPÓSIO DE APLICAÇÕES DE ÓPTICA E LASERS, 2017, São José dos Campos, Brasil. **Anais...** São José dos Campos, 2017. 68, 69, 87

ASTIER, P. et al. The supernova legacy survey: measurement of ω , λ and w from the first year data set. **Astronomy & Astrophysics**, v. 447, n. 1, p. 31–48, 2006. Disponível em:
<<https://doi.org/10.1051/0004-6361:20054185>>. 29

BARBON, R.; BUONDI, V.; CAPPELLARO, E.; TURATTO, M. Asiago Supernova catalogue. **VizieR Online Data Catalog**, v. 1, nov. 2009. 29

BLONDIN, S.; MANDEL, K. S.; KIRSHNER, R. P. Do spectra improve distance measurements of type ia supernovae? **Astronomy & Astrophysics**, v. 526, n. A81, February 2011. 28, 29, 43

BLONDIN, S. et al. The spectroscopic diversity of type ia supernovae. **The Astronomical Journal**, v. 143, n. 5, p. 126, 2012. 29

BLONDIN, S.; TONRY, J. L. Determining the type, redshift, and age of a supernova spectrum. **The Astrophysical Journal**, v. 666, n. 2, 2007. 21, 75

BRITANNICA, E. **Redshift**. 2018. Disponível em:
<<https://www.britannica.com/science/redshift>>. Acesso em: 08 jan. 2019. 32

- CAMBRIDGE PHOTOMETRY CALIBRATION SERVER. **Cambridge photometry calibration Server**. 2019. Disponível em: <<http://gsaweb.ast.cam.ac.uk/followup/>>. Acesso em: 09 fev. 2019. 29
- CAMPBELL, H. et al. Photometric science alerts from Gaia. In: WOZNIAK, P. et al. (Ed.). **The third hot-wiring the transient universe workshop**. [S.l.: s.n.], 2014. p. 43–50. 29
- CARROLL, B. W.; OSTLIE, D. A. **An introduction to modern astrophysics**. San Francisco: Pearson Addison-Wesley, 2007. 1358 p. 5, 7
- CHAMBERS, K.; WERNER, S. The pan-starrs1 surveys. 01 2014. Disponível em: <https://arxiv.org/abs/1612.05560>. 29
- DAMINELI, A.; STEINER, J. **Fascínio do Universo**. São Paulo: Odysseus, 2010. 120 p. 5
- DEPARTAMENTO DE FÍSICA DA SOUTHERN METHODIST UNIVERSITY. 2018. Disponível em: <<http://http://www.physics.smu.edu/>>. Acesso em: 12 abr. 2018. 5
- DOMINGOS, P. **O algoritmo mestre**. São Paulo: Novatec, 2017. 11, 12
- FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, v. 35, n. 1, p. 25–30, 2006. 11, 14, 15
- FILIPPENKO, A. V. Optical spectra of supernovae. **Annual Review of Astronomy and Astrophysics**, v. 35, n. 1, p. 309–355, 1997. 8, 9
- FLEWELLING, H. et al. The pan-starrs1 database and data products. 12 2016. Disponível em: <https://arxiv.org/abs/1612.05243>. 29
- GAL-YAM, A. et al. The caltech core-collapse project (cccp). 10 2004. Disponível em: <http://www.astro.caltech.edu/avishay/cccp.html>. 29
- GILLOCHON, J.; PARRENT, J.; KELLEY, L. Z.; MARGUTTI, R. An open catalog for supernova data. **The Astrophysical Journal**, v. 835, n. 1, p. 64, 2017. 29, 61, 87
- HAMUY, M. et al. The Carnegie supernova project: the low-redshift survey. **The Astronomical Society of the Pacific**, v. 118, p. 2–20, jan. 2006. 29

- HARUTYUNYAN, A. **Automatic objective classification of supernovae**. Tese (Doutorado em Astronomia) — Università degli Studi de Padova, Pádua, 2008. 22
- HAYKIN, S. **Redes neurais: princípios e práticas**. Porto Alegre: Bookman, 2001. 11, 12, 13, 15, 17
- HOWERTON, S. C. **CRTS SNhunt: the first five years of supernova discoveries**. [S.l.: s.n.], 2017. 29
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning with applications in R**. New York: Springer, 2013. 440 p. 79
- KOVÁCS, Z. L. **Redes neurais artificiais fundamentos e aplicações**. São Paulo: Livraria da Física, 2006. 11, 12
- LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977. 38
- LATEST SUPERNOVAE. **Latest supernovae**. 2019. Disponível em: <<http://www.rochesterastronomy.org/supernova.html>>. Acesso em: 09 fev. 2019. 29
- LUGER, G. F. **Inteligência artificial**. São Paulo: Pearson, 2013. 15, 16
- MAGNIER, E. et al. Pan-starrs data processing system. 12 2016. Disponível em: <https://panstarrs.stsci.edu/>. 29
- MATHESON, T. et al. Optical spectroscopy of type ia supernovae. **The Astronomical Journal**, v. 135, n. 4, 2008. 29
- MENCIA, E. L.; FURNKRANZ, J. Pairwise learning of multilabel classifications with perceptrons. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2008, Hong Kong, China. **Proceedings...** Hong Kong, 2008. p. 2899–2906. ISSN 2161-4393. 18, 19
- MODJAZ, M. et al. Optical spectra of 73 stripped-envelope core-collapse supernovae. **The Astronomical Journal**, v. 147, n. 5, 2014. 8, 9, 29
- MÓDOLO, M. **Classificação automática de supernovas usando redes neurais artificiais**. Tese (Doutorado em Computação Aplicada) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2016. Disponível

em: <<http://mtc-m21b.sid.inpe.br/col/sid.inpe.br/mtc-m21b/2016/04.20.20.44/doc/publicacao.pdf>>. xi, xiii, 1, 23, 24, 41, 42, 43, 66, 72, 87

NASCIMENTO, F. J. B.; ARANTES FILHO, L. R.; GUIMARÃES, L. N. F. Intelligent classification of supernovae using artificial neural networks.

Inteligencia Artificial, v. 22, n. 63, p. 39–60, 2019. DOI:

<https://doi.org/10.4114/intartif.vol22iss63pp39-60>. 41

OLIVEIRA FILHO, K. S.; SARAIVA, M. F. O. **Astronomia e astrofísica**. Rio Grande do Sul: UFRGS, 2014. 784 p. 1, 6, 7

OONG, T. H.; ISA, N. A. M. One-against-all ensemble for multiclass pattern classification. **Applied Soft Computing**, v. 12, n. 4, p. 1303 – 1308, 2012. ISSN 1568-4946. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1568494611004789>>. 17

OSÓRIO, F. S. Redes neurais artificiais: do aprendizado natural ao aprendizado artificial. In: FÓRUM E SEMINÁRIO DE INTELIGÊNCIA ARTIFICIAL DA ULBRA / SBC. **Anais...** Canoas, 1999. p. 1–32. 84

PATAT, F. et al. The type ia supernova 1994d in ngc 4526: the early phases. **Monthly Notices of the Royal Astronomical Society**, v. 278, p. 111–124, 12 1995. 76

PERLMUTTER, S. et al. Measurements of omega and lambda from 42 high-redshift supernovae. **The Astrophysical Journal**, v. 517, n. 2, p. 565, 1999. 1

PIMENTA, E. M. C. **Abordagens para decomposição de problemas multi-classe: os códigos de correção de erros de saída**. 102 p. Dissertação (Mestrado em Informática) — Faculdade de Ciências da Universidade do Porto, Porto, 2004. 18, 19

QUILES, M. G. **Sistema de visão baseado em redes neurais artificiais para o controle de robôs móveis**. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) — Universidade de São Paulo (USP), São Paulo, 2004. 14

RICHARDSON, D. et al. SUSPECT, The online supernova spectrum archive: year two. v. 34, p. 1205, 12 2002. Disponível em:

<http://adsabs.harvard.edu/abs/2002AAS...201.5609R>. 29

RIESS, A. G. et al. Time dilation from spectral feature age measurements of type ia supernovae. **The Astronomical Journal**, v. 114, 07 1997. 75, 77

RIESS, A. G. et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. **The Astronomical Journal**, v. 116, n. 3, p. 1009, 1998. 1

SASDELLI, M. et al. Exploring the spectroscopic diversity of type ia supernovae with dracula: a machine learning approach. **Monthly Notices of the Royal Astronomical Society**, v. 461, n. 2, p. 2044–2059, 2016. 24, 25

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627–1639, 1964. Disponível em: <<https://doi.org/10.1021/ac60214a047>>. 32

SILVERMAN, J. M.; KONG, J. J.; FILIPPENKO, A. V. Berkeley supernova ia program – ii. initial analysis of spectra obtained near maximum brightness. **Monthly Notices of the Royal Astronomical Society**, v. 425, n. 3, p. 1819–1888, 2012. 26

SILVERMAN, J. M. et al. Berkeley supernova Ia program - I. observations, data reduction and spectroscopic sample of 582 low-redshift type Ia supernovae. **Monthly Notices of the Royal Astronomical Society**, v. 425, p. 1789–1818, set. 2012. 29

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing and Management**, v. 45, p. 427–437, 2009. 37, 39

STERNBERG ASTRONOMICAL INSTITUTE SUPERNOVA CATALOGUE. **SAI supernovae catalog**. 2019. Disponível em: <<http://www.sai.msu.su/sn/sncat/>>. Acesso em: 11 fev. 2019. 29

SUN, F.; GAL-YAM, A. Quantitative classification of type i supernovae using spectroscopic features at maximum brightness. **ArXiv e-prints**, 2017. Disponível em: <<https://arxiv.org/abs/1707.02543>>. 25

TENENBAUM, J. B.; SILVA, V. d.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, n. 5500, p. 2319–2323, 2000. Disponível em: <<http://science.sciencemag.org/content/290/5500/2319>>. 60

- TURATTO, M. Classification of supernovae. In: K.W., W. (Ed.). **Supernovae and gamma-ray bursters**. Berlin: Springer, 2003. p. 21–36. 1, 7, 27
- TURATTO, M.; BENETTI, S.; PASTORELLO, A. Supernova classes and subclasses. **AIP Conference Proceedings**, v. 937, n. 1, p. 187–197, 2007. 8
- WANG, X. et al. Improved distances to type ia supernovae with two spectroscopic subclasses. **The Astrophysical Journal Letters**, v. 699, n. 2, p. L139, 2009. xv, 24, 25
- WIKIMEDIA. **Star life cycle chart**. 2017. Disponível em: <https://commons.wikimedia.org/wiki/File:Star_Life_Cycle_Chart.jpg>. Acesso em: 09 fev. 2019. 6
- WOOSLEY, S. E.; WEAVER, T. A. The physics of supernova explosions. **Annual Review of Astronomy and Astrophysics**, v. 24, n. 1, p. 205–253, 1986. 6
- WYRZYKOWSKI, L. et al. Ogle-iv real-time transient search. **Acta Astronomica**, v. 64, 09 2014. 29
- YARON, O.; GAL-YAM, A. WISeREP an interactive supernova data repository. **The Astronomical Society of the Pacific**, v. 124, p. 668, jul. 2012. 29
- ZHU, Y. et al. Kunlun dark universe survey telescope. In: THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING. **Proceedings...** Bellingham, 2014. v. 9145, p. 9145. 2

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Contam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.