Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

RING-Id: Uma ferramenta robusta para a identificação automática de galáxias peculiares do tipo aneladas

Elinavilmo de Morgado Santos

Feira de Santana

2018

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Elinavilmo de Morgado Santos

RING-Id: Uma ferramenta robusta para a identificação automática de galáxias peculiares do tipo aneladas

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

Orientador: Prof. Dr. Eduardo Brescansin de Amôres

Coorientador: Prof. Dr. Maximiliano Luiz Faúndez-Abans

Feira de Santana

2018

Ficha Catalográfica - Biblioteca Central Julieta Carteado - UEFS

S234r Santos, Elinavilmo de Morgado
RING-Id: uma ferramenta robusta para a identificação automática de galáxias peculiares do tipo aneladas / Elinavilmo de Morgado Santos. - 2018.
93f.: il.

Orientador: Eduardo Brescansin de Amôres.
Coorientador: Maximiliano Luiz Faúndez-Abans.
Dissertação (mestrado) - Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2018.

1. RING-Id (Software) – Levantamento astronômico. 2. Galáxias peculiares aneladas - Características. 3. Redes Neurais Artificiais. 4. Processamento de imagens (computação). I. Amôres, Eduardo Brescansin de, orient. II. Faúndez-Abans, Maximiliano Luiz, coorient. III. Universidade Estadual de Feira de Santana. IV. Título.

CDU: 004:524.6

Solange dos Santos Rocha – Bibliotecária CRB5/1639

Elinavilmo de Morgado Santos

RING-Id: Uma ferramenta robusta para a identificação automática de galáxias peculiares do tipo aneladas

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

Feira de Santana, 09 de agosto de 2018

BANCA EXAMINADORA

Eduardo Brescansin de Amôres

Dr. Eduardo Brescansin de Amôres (Orientador)
Universidade Estadual de Feira de Santana

Mirco Ragni

Dr. Mirco Ragni
Universidade Estadual de Feira de Santana

Ronaldo Savarino Levenhagen

Dr. Ronaldo Savarino Levenhagen
Universidade Federal de São Paulo

Abstract

In this Dissertation, is presented a robust tool to identify peculiar ring galaxies automatically, called ($RING-Id$) that uses Neural Artificial Net and the $SURF$ Algorithm by means the features extraction and the $LIRe$. As a data set, was used the catalogues of Faúndez-Abans & Oliveira-Abans (1998) and Moiseev et al. (2011) and DSS images. Due to the quality of those images and to have a good statistical sample of galaxies, at the moment the tool identify two categories of ring galaxies, the elliptical and the polar ring. The training and classification samples have 124 and 53 images, respectively. In the training phase, the accuracy and the F-Score were 98,0% and 98,6%, respectively. On the other hand, for the classification phase, those values were 81,0% and 73,0%, respectively. The tool contains two modules: i-) training, ii-) classification and visualisation. It can also be adapted to be used in any Large Astronomical Survey image to identify any other peculiar galaxy category.

Keywords: peculiar ring galaxies, large astronomical surveys, pattern recognition, image processing.

Resumo

A presente Dissertação aborda a elaboração de uma ferramenta robusta para a identificação automática de galáxias peculiares aneladas, por meio da extração de características com o uso do Algoritmo *SURF*, da biblioteca (*LIRe*) e da elaboração de um software (*RING−Id*) baseado em Redes Neurais Artificiais. Foi utilizada como base de dados, os catálogos de galáxias aneladas, de Faúndez-Abans e Oliveira-Abans (1998) e Moiseev et al. (2011). Devido a aspectos de qualidade das imagens e ao número representativo de galáxias para as categorias de galáxias aneladas, atualmente, a ferramenta identifica de forma automática, as categorias de aneladas elípticas e polares. As amostras de treinamento e de classificação, contém 124 e 53 imagens, respectivamente, para ambas as categorias. Na fase de treinamento, a taxa de acerto utilizando-se da métrica de acurácia foi de 98,0% e da Medida-F de 98,6%. Para a fase de classificação, a taxa de acurácia foi de 81,0% e da Medida-F 73,0%. O software contém os módulos de treinamento, classificação e de visualização de dados, e pode ser facilmente adaptado para o uso em imagens de qualquer Grande Levantamento Astronômico, para a identificação automática de demais categorias de aneladas ou de outras categorias de galáxias peculiares, ou objetos astronômicos.

Palavras-chave: galáxias peculiares, grandes levantamentos astronômicos, reconhecimento de padrões, processamento de imagens.

Prefácio

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

A dissertação foi desenvolvido dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. Eduardo Brescansin de Amôres, e como co-orientador o Dr. Maximiliano Luiz Faúndez-Abans.

Agradecimentos

Agradeço primeiramente a Deus, pois, nos acompanha em nossos momentos difíceis; foi o meu guia em todos os instantes da minha vida.

Agradeço aos meus pais e irmãos que sempre acreditaram em meus estudos.

A minha filha Stephanie Lopes Morgado Santos, com seus 7 anos, meu orgulho de ser pai e incentivo em meus pensamentos.

A minha família, amigos e professores, os meus sinceros agradecimentos por terem me dedicado parte especial das suas vidas.

Ao meu orientador, Prof. Dr. Eduardo Brescansin de Amôres e ao meu coorientador Prof. Dr. Maximiliano Luiz Faúndez-Abans. Agradeço à ambos pela dedicação e rigor científico, sendo que em todos os momentos me acompanharam neste trabalho, com paciência e compreensão. Sem eles, a Dissertação não teria sido possível. Agradeço também ao Dr. Faúndez-Abans pela confiança e pela possibilidade, de trabalhar sob a sua coorientação em um tema no qual, ele é um dos especialistas em âmbito mundial.

Agradeço a equipe do LAi/INCT-A por possibilitar a utilização de seu cluster: GINA/Alphacrucis. O que veio a permitir a conclusão deste trabalho em tempo hábil.

Agradeço a Banca Examinadora, Prof. Dr. Mirco Ragni e ao Prof. Dr. Ronaldo Savarino Levenhagen pelas contribuições ao meu trabalho.

Agradeço ao meu amigo e enquanto colega de trabalho Jairo Henrique dos Santos Calmon pelas revisões algumas vezes deste trabalho e incentivo, o meu muito obrigado.

Agradeço aos Prof. Dr. Angelo Amâncio Duarte e ao Prof. Dr. Iranderly Fernandes de Fernandes enquanto coordenadores do Programa de Pós-Graduação em Computação Aplicada, pela disponibilidade, atenção dispensada e profissionalismo.

Agradeço ao Prof. Dr. Angelo Conrado Loula e ao Prof. Dr. Rodrigo Tripodi Calumby pelas observações e correções no Exame de Qualificação. Ao Prof. Dr. Angelo Conrado Loula também gostaria de agradecer pelas aulas no PGCA onde fui um dos seus alunos nas disciplinas de Inteligência Computacional e Mineração de

Dados. Em especial ao Prof. Dr. Rodrigo Tripodi Calumby pelas discussões sobre a biblioteca *LIRe* e pelo algoritmo *SURF* utilizados nesta Dissertação.

Agradeço a todos o servidores do UEFS, em especial a Patricia Tavares Santos enquanto lotada na Secretária do Colegiado de Pós-Graduação em Computação Aplicada pela sua cortesia e profissionalismo.

Agradeço aos colegas de trabalho do Instituto Federal de Educação, Ciência e Tecnologia Baiano, em especial aos colegas do *Campus* Alagoinhas e Catu pela compreensão e incentivo.

Gostaria de agradecer também a todos que contribuíram, direta ou indiretamente, na minha formação, para todos vocês o meu muito obrigado.

Este trabalho utilizou os equipamentos do Laboratório de Astroinformática (IAG/USP, NAT/Unicsul), cuja aquisição foi possibilitada pela agência brasileira FAPESP (processo 2009/54006-4) e pelo INCT-A.

Sumário

Lista de Publicações

”Detecção automática de Galáxias Peculiares do tipo Aneladas”; apresentado e publicado nos anais do III WPOS/ERBASE 2016, ISSN 2177-4692. Autores: Santos, E. M.; Amôres, E. B.; Faúndez-Abans, M.A.; Oliveira-Abans, M.; Oliveira-Abans, M.;da Rocha-Poppe, P.C.; Martins, F.V.A.

Lista de Tabelas

Lista de Figuras

Lista de Abreviações

Abreviação	**Descrição**
AM87	Arp & Madore, 1987 (AM87)
BoF	*Bag-of-Features*
CEDD	*Color and Edge Directivity Descriptor*
CCD	*Charge-compled device*
E	*Elliptical galaxy*
FCTH	*Fuzzy Color Texture Histogram*
Gaia	*Global Astrometric Interferometer for Astrophysics*
Irr	*Irregular galaxy*
IDL	*Interactive Data Language*
LSST	*Large Synoptic Survey Telescope*
LIRe	*Lucena Image Retrieval*
P	*Polar galaxy*
PRGs	*Polar Ring Galaxies*
RNA	*Artificial Neural Network*
SDSS	*Sloan Digital Sky Survey*
SURF	*Speeded Up Robust Features*
SIFT	*Scale Invariant Feature Transform*
VVV	*VISTA Variables in the Via Láctea*
2MASS	*Two Microns All Sky Survey*

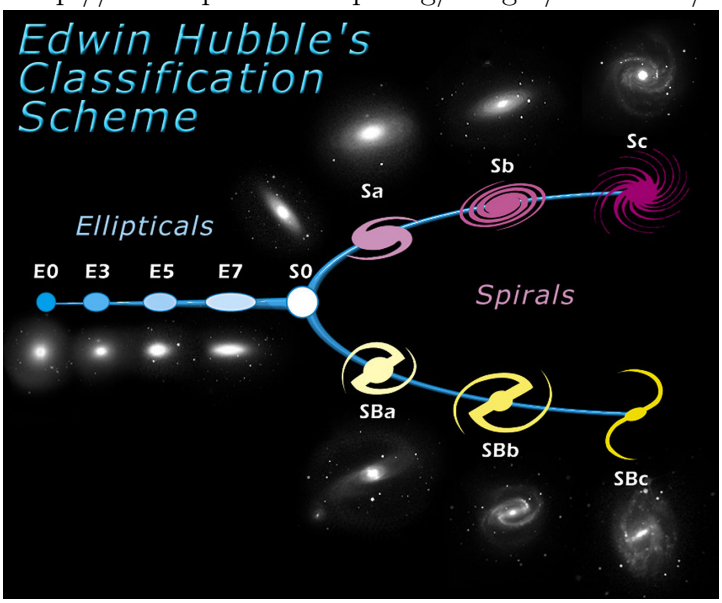
Capítulo 1

Introdução

Uma das principais motivações para a identificação e classificação de galáxias, reside no fato de que é possível separar distintos tipos de galáxias, que tenham características morfológicas semelhantes, para fins de estudo e conhecimento dos processos relacionados à sua formação e evolução.

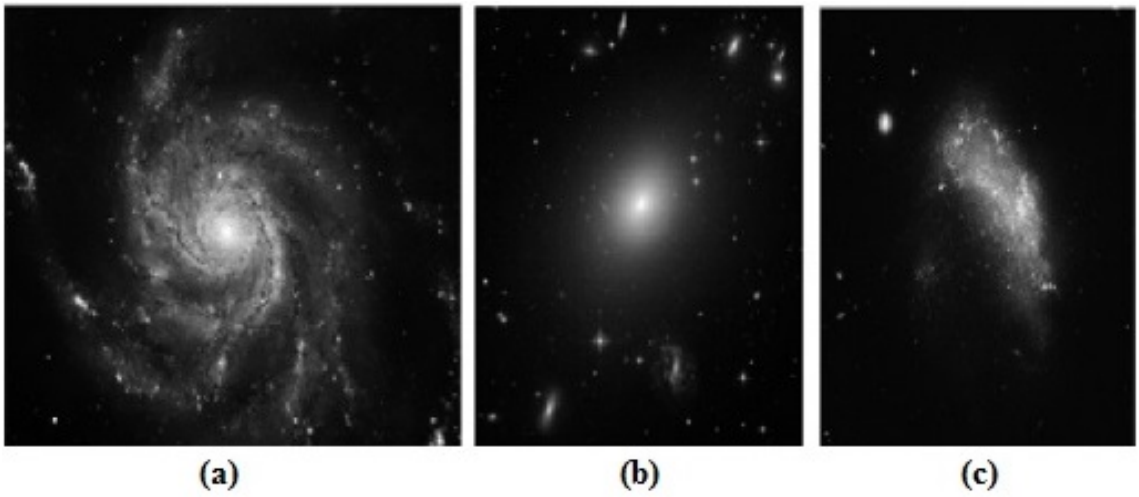
Em 1926, Edwin Hubble publicou um artigo sobre a classificação de galáxias. No trabalho, a classificação das galáxias foi baseada em sua estrutura e morfologia. Os principais tipos morfológicos das galáxias são representados no Esquema de *Hubble* (Figura 1.1). Nesse esquema, as galáxias são separadas basicamente em três classes: Elípticas (E), Espirais (S) e Irregulares (Irr).

Figura 1.1: Diagrama de *Hubble*. Fonte: Adptado de http://www.spacetelescope.org/images/heic9902o/



De acordo com Ianishi et al. (2017) uma galáxia do tipo elíptica tem aparência esférica achatada, e possui uma distribuição suave de luz. As galáxias espirais apresentam uma estrutura com braços espirais que contém basicamente, gás, poeira e estrelas (Lépine et al., 2001; Amôres & Lépine, 2005; entre outros). As galáxias irregulares não possuem um padrão ou forma definida, as mais conhecidas são a Pequena e a Grande Nuvem de Magalhães. A Figura 1.2 apresenta exemplos representativos desses tipos de galáxias.

Figura 1.2: Exemplos de galáxias: Espiral (a), Elíptica (b) e Irregular (c). Fonte: Adaptado de Ianishi et al. (2017).



Segundo Naim & Lahav (1997), as galáxias que não se enquadram no Esquema de *Hubble* podem ser enquadradas como galáxias peculiares. Essa classificação foi motivada, devido ao enorme incremento de catálogos a partir da análise de galáxias. A natureza das galáxias peculiares depende da definição utilizada. Ainda segundo os autores, as galáxias peculiares correspondem à aproximadamente 0.1% da quantidade total de galáxias existentes no Universo.

Um dos pioneiros do estudo de galáxias peculiares foram Halton Christian Arp e Barry Madore que publicaram em 1987, um trabalho intitulado "Atlas de Galáxias Peculiares" o que hoje é uma base de estudos amplamente utilizada pelos astrônomos, que objetiva o entendimento da evolução e morfologia das galáxias. O Catálogo de Arp & Madore (1987, daqui por diante, AM87) contém aproximadamente 6.000 galáxias peculiares distribuídas por 25 categorias.

O Catálogo de AM87 foi inédito não tendo como base nenhum trabalho anterior, sendo desenvolvido de maneira empírica. O catálogo obteve grande sucesso, o que refletiu em grandes contribuições para a área da Astrofísica extra-galáctica, graças ao advento de telescópios com maior capacidade de observação, a exemplo do Telescópio *Schmidt*, que deu uma nova perspectiva, indisponível para as primeiras gerações de "astrônomos extra-galácticos".

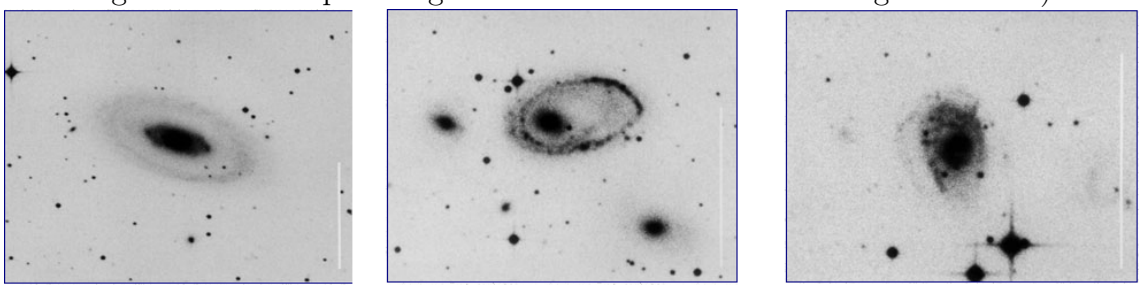
Dentre os diversos tipos de galáxias catalogadas em AM87, existem aquelas que possuem um anel bem definido em seu contorno, sendo chamadas de galáxias aneladas. Para explicar a formação desses tipos de objetos, existem algumas abordagens que envolvem a colisão de uma nuvem ou galáxia com outra galáxia de uma maneira especial, geralmente através do centro da galáxia, que se torna a galáxia anelada (AM87). As galáxias aneladas representam a Categoria 6 do Catálogo de AM87, e representam uma fração de 3.1% do total das galáxias peculiares, as quais são os objeto de estudo da Dissertação.

1.1 As galáxias aneladas

Arp & Madore (1987) no *Catalogue of Southern Peculiar Galaxies and Associations* classificam como objetos da Categoria 6 do catálogo, qualquer tipo de galáxia que contém um anel luminoso em seu contorno. Nessa categoria temos três subcategorias de galáxias aneladas nesse catálogo: *Displaced Nuclei*, *Centered Nuclei* e *Irregular Rings*. A Figura 1.3 apresenta exemplos de galáxias aneladas do Catálogo de AM87.

De acordo com Whitmore (1990) existem duas possibilidades de galáxias aneladas, as normais e as peculiares. As galáxias aneladas normais geralmente são galáxias espirais barradas com anel interno. A formação do anel se deve à ressonância da barra em alta velocidade em relação as estrelas e gás. Por sua vez, as galáxias aneladas peculiares são baseadas em um processo de interação com outras galáxias, e são bem mais raras.

Figura 1.3: Exemplos de galáxias aneladas. Fonte: Catálogo de AM87).

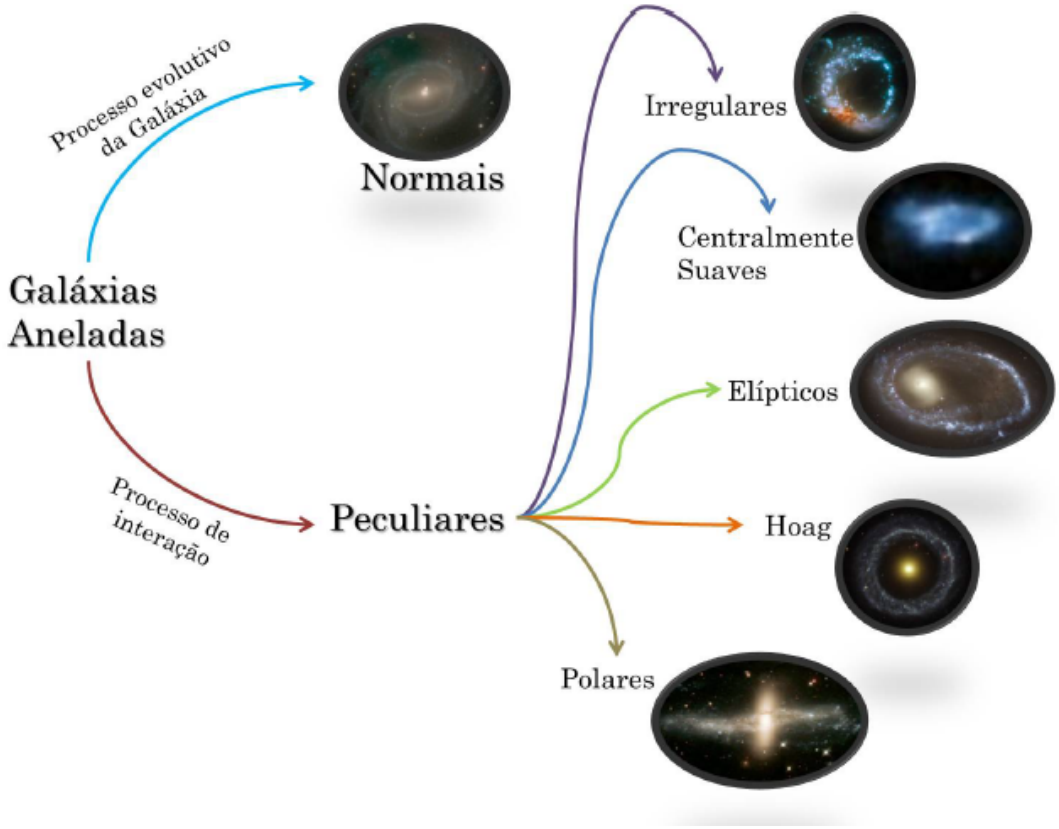


Few & Madore (1986) propuseram uma simples divisão das galáxias aneladas do Hemisfério Sul em dois grandes sub-tipos, anéis do Tipo-P e Tipo-O, com o núcleo formado por estruturas provavelmente devido a colisões e com companheiras próximas e com um núcleo suave central, sem companheira evidente.

Faúndez-Abans e Oliveira-Abans (1998a) citam alguns trabalhos pioneiros em estudar a estrutura de algumas galáxias, por exemplo, a investigação para reforçar estruturas das galáxias tendo como uso, técnicas de processamento de imagens por

meio da Transformadas de Fourier. Para os autores, as técnicas de processamento de imagem podem ser uma das ferramentas mais importantes no processamento de grandes conjuntos de dados, encontrando possíveis dados ocultos no sinal da galáxia, tornando os resultados mais adequados para a classificação do que os dados originais obtidos por meio de imagens bidimensionais.

Figura 1.4: Exemplos de tipos de Galáxias Aneladas. Fonte: Freitas-Lemes (2014).



Uma breve explicação de cada um dos objetos apresentados na Figura 1.4 de acordo com Freitas-Lemes (2014) são sucintamente descritos a seguir:

1. Anéis de *Hoag*: as galáxias aneladas desse tipo contém um anel circular em torno de um pequeno bojo;
2. Anéis Polares: são formadas por uma galáxia hospedeira rodeada de um anel de gás, poeira e estrelas que orbita um plano aproximadamente polar, em relação ao plano principal da galáxia hospedeira;
3. Anéis Elípticos: as elípticas possuem várias subdivisões. Os anéis estão frequentemente presentes nestas estruturas, mostram pertubações em sua estrutura;

4. Anéis Centralmente Suaves: Esses objetos apresentam anéis com nódulos e o interior a este aparece "vazio", não apresentam estrutura como bojo ou núcleo;
5. Anéis Irregulares: uma boa quantidade de galáxias mostram distorções irregulares em suas estruturas que se assemelham a pseudo-anéis, braços e caudas curvas.

Figura 1.5: Exemplos de boas candidatas, segundo os autores, à Polar *Ring*. Fonte: Adaptado de Moiseev et al (2011).

A componente difusa da componente estelar de galáxias com formato de anéis, como classificadora foi proposta originalmente por Faúndez-Abans et al. (1992). Faúndez-Abans & Oliveira-Abans (1998a) propuseram cinco divisões para galáxias peculiares de acordo com sua morfologia e bojo, baseados em inspeção visual de 489 objetos.

Atualmente essa é a melhor classificação apresentada. As cinco subcategorias propostas pelos autores são também descritas na Figura 1.4. As cinco subcategorias são: Polar (Pa,b,c), *Hoag* (HLa,b), Elíptica (Ea,b,c), Irregular (I) e Anéis Centralmente Suaves (CS). Outros trabalhos discutindo a classificação e aparência de galáxias peculiares aneladas são descritos em Faúndez-Abans & Oliveira-Abans (1998b) e Faúndez-Abans et al. (1994). Em relação às galáxias aneladas polares, o primeiro estudo com o objetivo de identificar galáxias candidatas a terem um anel polar as (*Polar Ring Galaxies*) foi realizado por Whitmore (1990). A sua compilação em forma de Atlas fotográfico de *Polar Ring Galaxies* (PRGs) e objetos relacionados, apresentam um total de 157 objetos.

Projetos como o *Galaxy Zoo* oferecem uma boa oportunidade para encontrar novas candidatas à *PRGs*, pois não existe uma classificação única para as *PGRs*, apenas prováveis candidatas (Finkelman et al., 2012).

Moiseev et al. (2011) compilaram um catálogo de galáxias aneladas polares tendo como base os dados do *SDSS*, assim como uma classificação prévia desses objetos no projeto *Galaxy Zoo*, sendo que de uma amostra inicial de aproximadamente 40.000 galáxias foram selecionadas 275 como possíveis *PRG*.

Os autores também realizaram observações do espectro de algumas das galáxias, com o intuito de verificar se de fato possuíam um anel polar, o qual foi identificado para cinco galáxias. Alguns objetos do Catálogo de Moiseev et al. (2011) são apresentados na Figura 1.5.

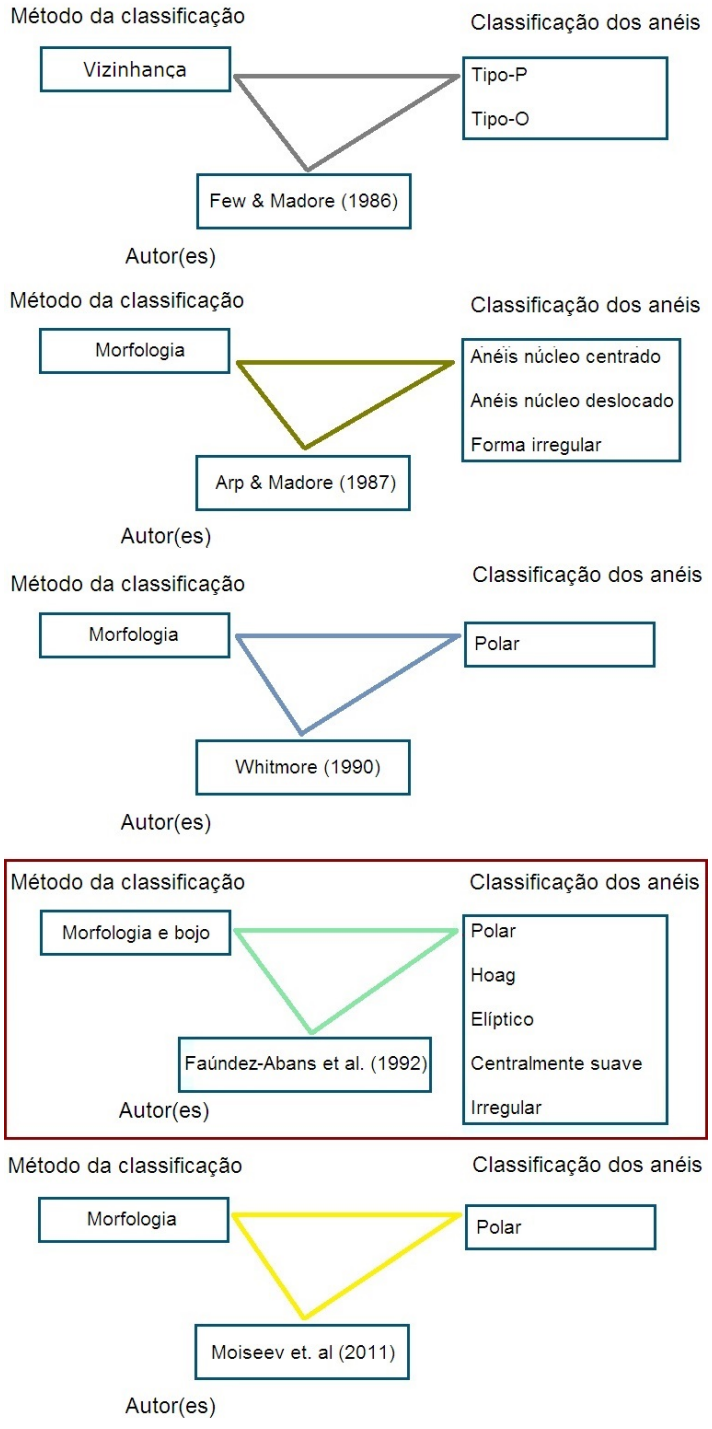
Cabe ressaltar que até o início da década, existiam apenas três galáxias para as quais o movimento de estrelas no anel polar foi estudado, NGC 4650A (Swaters & Rubin, 2003), UGC 5119 (Merkulova et al., 2008) e UGC 2748 (Merkulova et al., 2008).

Segundo Freitas-Lemes et al. (2012) as galáxias com anéis polares se formam durante um evento secundário em torno de uma galáxia preexistente, como exemplo do acréscimo de gás e estrelas pela galáxia hospedeira. Existem três estruturas básicas conforme menciona Faúndez-Abans e Oliveira-Abans (1998b) que podem ser associados a esta família:

1. *Spindle*: são galáxias anel polar tradicionais. Esses objetos apresentam uma galáxia central (a galáxia S0) com um anel quase perpendicular ao eixo maior galáctico (anel polar).
2. *Saturn*: esses objetos têm uma protuberância esférica com um anel rodeada brilhante.
3. *Worm-like*: Whitmore et al. (1990) sugerem que esse tipo de objeto são possíveis candidatos a galáxias com anéis polares. Essa categoria engloba os objetos que tenham uma galáxia hospedeira alongada ou objetos que possuem o anel e o bojo.

Um resumo para a classificação de alguns autores é apresentado na Figura 1.6, a qual apresenta representações em forma de triângulos, onde suas arestas contém informações de autores, método da classificação e classes dos anéis. O retângulo com bordas vermelhas presente na quarta imagem representa a melhor classificação dos anéis de acordo com Faúndez-Abans et al. (1992).

Figura 1.6: Formas de anéis classificados de acordo com os autores.

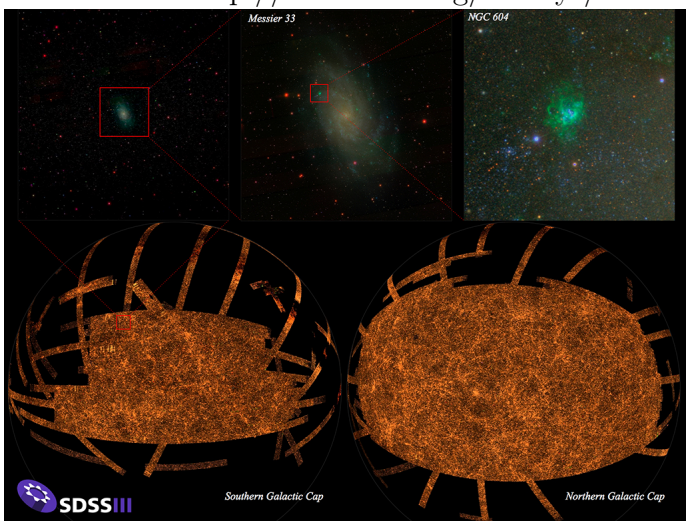


1.2 O reconhecimento de padrões em grandes volumes de dados astronômicos

Atualmente a Astronomia tem produzido grandes volumes de dados de observações de extensas regiões do céu (ou a sua totalidade). Esse tipo de observações são chamados de Grandes Levantamentos Astronômicos, os quais têm por objetivo mapear grandes áreas do céu.

Dentre os grandes levantamentos astronômicos, podemos citar alguns, tais como: o 2MASS[1] (Skrutskie et al., 2006), o SDSS[2] (Abazajian et al., 2009), o VVV[3] (Saito et al., 2012; Amôres et al., 2012; Amôres et al., 2013), o Gaia[4] (Brown et al., 2018), dentre outros. Essas observações compreendem um grande volume de dados. A Figura 1.7 no painel superior apresenta o exemplo de algumas imagens observadas pelo *SDSS*, no caso uma galáxia e uma região de formação estelar observadas pelo *SDSS*. O painel inferior apresenta a cobertura do *SDSS* tanto no Hemisfério Norte quanto no Hemisfério Sul, respectivamente.

Figura 1.7: Painel superior: exemplos de imagens em falsa cor obtidas pelo SDSS; painel inferior: mapa da cobertura das regiões observadas.
Fonte: http://www.sdss.org/surveys/



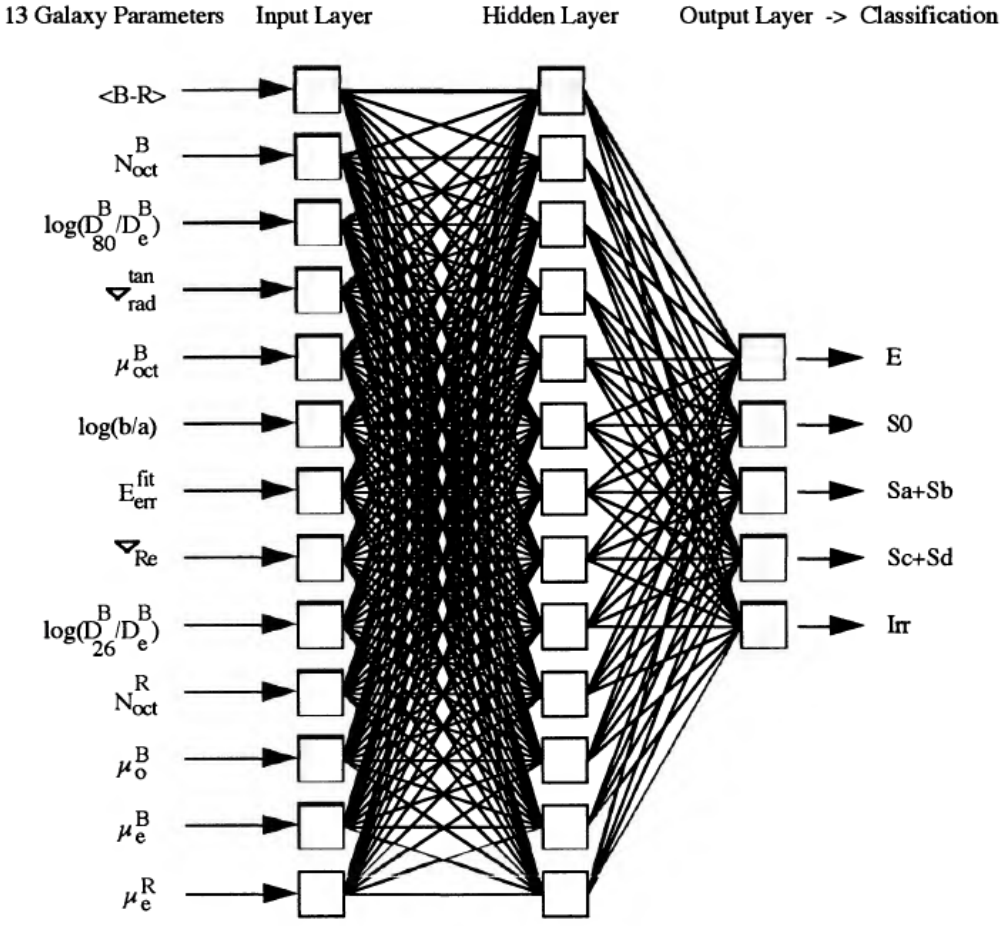
Devido a essa imensa quantidade de dados observados é importante procurar uma forma de classificar automaticamente os objetos, em nosso caso, as galáxias aneladas peculiares. As Redes Neurais Artificiais (RNAs) podem ser aplicadas para a classificação desses dados, pois são algoritmos de computador, que tem a possibilidade de

[1] *Two Microns All Sky Survey*
[2] *Sloan Digital Sky Survey*
[3] *VISTA Variables in the Via Láctea*
[4] *Global Astrometric Interferometer for Astrophysics*

prever um quadro conveniente para a classificação de objetos. Elas também estão relacionadas à métodos estatísticos utilizados na Astronomia, e em outros campos, como por exemplo, os Métodos Bayesianos e a Análise de Componente Principal (Lahav, 1996).

No trabalho de Storrie-Lombardi et al. (1992) é apresentado um exemplo de uma RNA que usa alguns parâmetros com informações referentes a brilho, grau de elipticidade, assimetria da galáxia, diâmetro, dentre outros, que foram utilizados como entradas da rede para fins de classificação para alguns diferentes tipos de galáxias. A estrutura da rede utilizada pelo autor é apresentada na Figura 1.8.

Figura 1.8: Exemplo de uma RNA utilizada por Storrie-Lombardi et al. (1992).



A configuração da rede neural utilizada no estudo de Storrie-Lombardi et al. (1992) consistiu em 13 entradas, 5 possibilidade de saídas (E, SO, $Sa{+}Sb$, $Sc{+}Sd$, Irr) e entre as camadas de entrada e saída, uma camada oculta com 13 neurônios. Para treinamento da rede foi utilizado o algoritmo *backpropagation*.

Considerando os resultados obtidos pela RNA da classificação de 5.217 galáxias, Storrie-Lombardi et al. (1992) notaram que 64% da rede estava em concordância a

classificação feita por inspeção visual e 54% com os resultados obtidos por meio de uma outra técnica de classificação, que usou um classificador linear estatístico.

Storrie-Lombardi et al. (1992) apontaram que as RNAs apresentaram uma significante melhoria conforme apresentado na Figura 1.9, que apresenta a comparação entre a análise automática feita pela RNA (esquerda) e a comparação feita por meio de inspeção visual (direita).

Figura 1.9: Classificação das galáxias: esquerda (RNA), direita (inspeção visual). Fonte: Storrie-Lombardi et al. (1992).

	(a) ANN					(b) ESO AUTO				
Class	E	S0	Sa+Sb	Sc+Sd	Irr	E	S0	Sa+Sb	Sc+Sd	Irr
E	203	77	25	1	5	197	87	17	5	5
S0	109	229	240	7	2	184	218	155	28	2
Sa+Sb	12	85	1281	218	15	106	12	791	664	38
Sc+Sd	1	4	304	415	36	22	11	24	631	72
Irr	0	0	53	69	126	22	9	31	42	144

Ball (2001) aplicou as técnicas de RNAs para classificar tipos de galáxias. O trabalho usou um catálogo originado dos dados do SDSS. Em outro estudo usando os dados classificados pelo projeto *Galaxy Zoo*, Banerji et al. (2008) utilizaram uma rede neural artificial, que fez uso de um subconjunto de treino das amostras classificadas pelo olho humano. Na fase de testes, a rede buscou por meio do algoritmo de aprendizado de máquina uma reprodução dos objetos restantes da amostra conforme classificados pelos humanos. Segundo os autores, a RNA, teve resultados satisfatórios comparados com a classificação humana.

Projetos como o desenvolvido pela equipe do *Galaxy Zoo*[5] permitem envolver o público na classificação de galáxias. Tais classificações possibilitam que os astrofísicos, procurem ligações entre a aparência das galáxias, e as suas propriedades internas e externas. A morfologia de uma galáxia é difícil de quantificar de forma concisa.

A base do *Galaxy Zoo* é uma excelente possibilidade para desenvolver, e testar análises de imagens e algoritmos de visão computacional, para classificações automáticas de galáxias. Existem outros trabalhos na reprodução da classificação usando aprendizado de máquinas, mas serão necessários melhores sistemas, quando se lida com os produtos de dados dos telescópios da próxima geração (Kremer et al., 2017).

Shamir e Wallin (2014) apresentam uma ferramenta computacional, chamada de *WndChrm* que também pode ser usada na identificação de galáxias peculiares em interação. Os autores utilizaram os dados do *SDSS* para um conjunto de aproximadamente 400.000 imagens de galáxias, extraindo algumas características, tais como: informações de textura, forma, cor, bordas e distribuição estatística da intensidade

[5]https://zoo4.galaxyzoo.org/

dos pixels. Nesse trabalho são apresentados 500 novos pares de galáxias observados no *SDSS*. Cabe ressaltar, que os autores não abordaram a identificação de galáxias aneladas.

Dieleman et al. (2014) fizeram uso das galáxias do *Galaxy Zoo* para treinar os diferentes tipos morfológicos oriundos de imagens de galáxias. O tipo de rede utilizada foi uma sub-classe de redes neurais chamada de redes neurais convolutivas, que são padrões de redes neurais com conectividade restritas entre algumas das camadas. Para os autores o modelo é altamente confiável pois os resultados das predições são atribuídos as diversas características de galaxias relacionadas a: barra, espiral, bojo, arredondamento etc.

Em sua Dissertação de Mestrado, Cerqueira (2016) fez a identificação de pares de galáxias interagentes das categorias 1 e 2 do Catálogo de AM87, usando o software *WndChrm* com os dados do 2MASS no infravermelho próximo. O autor identifica uma lista de aproximadamente 1.500 pares de galáxias.

Kim e Brunner (2017) utilizaram redes neurais profundas para automatizar, o processo de separação entre estrelas e galáxias, utilizando informação do pixel da imagem, diretamente sem quaisquer extração de características, obtendo uma precisão de 90% para prever corretamente os tipos morfológicos das galáxias.

Tuccillo et al. (2017) fizeram uso da capacidade de redes convolucionais profundas, para fornecer propriedades paramétricas, para imagens de galáxias observadas pelo Telescópio *Hubble*. Os autores compararam os seus resultados aos obtidos com o software *GALFIT*[6]. Na fase de treinamento, o software foi aproximadamente cinco vezes mais rápido de que o *GALFIT* para resultados pelo menos similares aos do *GALFIT*. Devido ao seu alto nível de abstração com pouco intervenção humana, os autores mencionam que a aprendizagem profunda parece ser uma abordagem promissora.

Shamir (2017) propôs um método que permite pesquisar automaticamente bancos de dados de imagens de galáxias, e identificar quais são morfologicamente similares à uma determinada consulta definida pelo usuário. Dessa forma, o pesquisador fornece uma imagem de uma galáxia de interesse, e o sistema de reconhecimento de padrões que utiliza algoritmo para calcular a similaridade, reenvia automaticamente uma lista de galáxias que são visualmente semelhante à galaxia alvo. O algoritmo usa um conjunto abrangente de descritores, permitindo-lhe suportar diferentes tipos de galáxias, não é limitado a um conjunto finito de morfologias conhecidas.

Timmis & Shamir (2017) desenvolveram um método de visão computacional para identificar galáxias aneladas utilizando dados fornecidos pelo projeto PanSTARRS[7]. Com os resultados obtidos da aplicação do método, em aproximadamente 3 milhões de galáxias, um catálogo de 185 candidatas a galáxias aneladas foi construído com base em inspeção visual.

[6] https://users.obs.carnegiescience.edu/peng/work/galfit/galfit.html

[7] *Panoramic Survey Telescope and Rapid Response System*

1.3 Justificativa e motivação

Devido ao grande volume de dados oriundos de observações astronômicas, a análise humana, como p.ex., a inspeção visual de imagens, algumas com alta resolução e com uma grande quantidade de objetos é inviável. Para identificar determinados tipos objetos, em nosso caso particular galáxias aneladas, a análise humana isolada é uma tarefa que demanda muito tempo e recursos humanos.

Conforme mencionado anteriormente, existem projetos como o *Galaxy Zoo*, o qual tem por objetivo geral permitir a identificação de tipos morfológicos de galáxias por meio da inspeção visual feita por "cidadãos-cientistas". Essa classificação apesar de contar com o envolvimento de um grande número de pessoas, alcançando até a centena de milhares, e dessa forma, uma boa estatística das classificação; não tem o detalhamento da identificação de tipos específicos de galáxias. Outro aspecto, é que o *Galaxy Zoo* por estar baseado nos dados do SDSS, contém classificações para o Hemisfério Norte.

A confiabilidade no que se refere à determinação das propriedades de alguns objetos, muitas vezes não é obtida com precisão, pois os softwares de redução de dados realizam processamento em grandes áreas do céu. A ocorrência desse aspecto, deve-se ao fato de que os softwares não conseguem trabalhar de maneira adequada com objetos com baixo brilho, ou, até mesmo, algumas características específicas que são apenas encontradas em pequenas frações de objetos.

Uma alternativa consiste na análise das imagens originais, e o emprego de técnicas de reconhecimento de padrões. Cabe ressaltar, que muitas vezes, os softwares de redução de dados, não determinam as propriedades específicas de objetos que se pretende estudar, classificando simplesmente como galáxia ou estrela. A Computação Aplicada na Astrofísica enfrenta grandes desafios, ao lidar com grandes volumes de informações oriundas de imagens disponíveis e, ao mesmo tempo, obter os padrões significativos entre milhões de dados.

Por outro lado, conforme foi visto nas duas últimas seções, não existem muitos trabalhos que visam a identificação automática de galáxias aneladas peculiares. Conforme mencionado anteriormente, a busca por esse tipo de objetos, por meio de uma inspeção visual não é viável, devido à imensa quantidade de dados disponíveis, e a busca por novas técnicas para realizar essa tarefa é imprescindível. Essas técnicas buscam a classificação automática destes objetos para um posterior estudo mais detalhado.

Tendo como base esse aspecto, foi usada uma técnica de aprendizagem de máquina para fins de classificação de algumas classes de objetos, em dados de levantamentos astronômicos fotométricos. O intuito é a utilização de uma RNA para fazer uso de informações obtidas, por meio dessa técnica para descrever imagens baseadas em histogramas de palavras visuais.

1.4 Objetivos

O objetivo principal e os específicos da Dissertação são descritos a seguir.

Objetivos principais:

- A identificação automática de galáxias aneladas por meio das técnicas de Redes Neurais Artificiais (RNAs);
- a elaboração do software *RING−Id* que usa uma RNA.

Objetivos específicos:

- A obtenção e manipulação de imagens astronômicas;
- a compilação de um catálogo e de uma base de imagens de galáxias aneladas;
- a utilização da biblioteca *LIRe* para a extração de descritores de características em imagens;
- a elaboração de histogramas de palavras visuais com o uso da técnica *Bag-of-Features*;
- uso da medida de distância para identificar similaridades entre objetos;
- incorporação da biblioteca *Encog* ao software *RING−Id*;
- o treinamento e a classificação da RNA;
- a utilização das medidas de avaliação de resultados (acurácia e *F-Score*).

1.5 Aspectos de originalidade da Dissertação

O software (*RING-Id*) elaborado na presente Dissertação, assim como a sua aplicação são inéditos na literatura. Não existem trabalhos de reconhecimento de padrão usando RNAs, e nem tampouco, algorítimos de detecção e descrição de pontos de interesse, na procura por galáxias aneladas peculiares. Por outro lado, a atual Dissertação é um esforço inicial no sentido de realizar a identificação automática de galáxias aneladas em dados de grandes levantamentos astronômicos.

Uma vez desenvolvida e testada a ferramenta para a identificação de galáxias aneladas, bastarão algumas adaptações para usá-la na identificação não somente de galáxias aneladas mas de outros tipos de galáxias peculiares. Esse vem sendo um esforço do Grupo de Astrofísica de Grandes Levantamentos[8] do Departamento de Física da UEFS, em outros trabalhos de Mestrado já foram elaborados métodos automáticos para a identificação de galáxias interagentes e Sistemas do Tipo-M51. O atual trabalho terá um grande impacto quando aplicado à imagens de grandes levantamentos, assim como na identificação e consequente aumento de galáxias peculiares como um todo.

1.6 Organização da Dissertação

No Capítulo 2 é feita uma Revisão da Literatura sobre redes neurais. O Conjunto de Dados usado na Dissertação é apresentado no Capítulo 3. No Capítulo 4 é apresentada a Metodologia, com uma breve explicação sobre a extração de atributos em imagens e do algoritmo de detecção e descrição de pontos de interesse usando o SURF (*Speeded Up Robust Features*), bem como o *LIRe* (*Lucene Image Retrieval*) e a técnica *BoF* (*Bag-of-Features*) para elaboração dos histogramas de palavras visuais e as medidas de similaridade, assim como uma aplicação da técnica para galáxias aneladas. No Capítulo 5 é apresentada a ferramenta *RING−Id*, a sua aplicação e os resultados obtidos. No Capítulo 6 são apresentadas as Conclusões e Perspectivas.

[8] http://dgp.cnpq.br/dgp/espelhogrupo/7677110842270686

Capítulo 2

Revisão da literatura sobre Redes Neurais Artificiais

Uma rede neural artificial é uma ferramenta de modelagem de dados poderosa, que é capaz de capturar e representar complexas relações de entrada e saída. A motivação para o desenvolvimento da tecnologia de rede neural resultou do desejo de desenvolver um sistema artificial, que poderia executar "tarefas inteligentes" semelhantes às realizadas pelo cérebro humano (Hayati & Shirvany, 2007).

Ainda segundo os autores, o verdadeiro poder e vantagem das redes neurais reside na sua capacidade para representar, tanto os relacionamentos não-lineares quanto os lineares, assim como em sua capacidade de aprender essas relações diretamente, tendo como base os dados que estão sendo modelados.

As redes neurais são motivadas pela incrível capacidade paralela de processamento de cérebros biológicos, notadamente o cérebro humano. A principal motivação consiste em recriar as habilidades do neurônio biológico em modelos artificiais. Em computação, as redes neurais são destinadas à resolução de problemas, tais como: otimização, aproximações de funções, classificação, previsão de séries temporais, dentre outras aplicações (Muñoz, 2009).

Os modelos baseados em redes neurais buscam enfatizar a capacidade do cérebro para se adaptar ao mundo em que está situado, modificando as relações entre os neurônios individuais. Em vez de representar o conhecimento em lógica explícita, eles o fazem implicitamente, com um estabelecimento de padrões de relacionamentos (Luger, 2005).

As RNAs possuem uma visão multidisciplinar que pode ser empregada em diferentes campos de aplicações, tais como: na área médica, classificação de padrões, reconhecimento de faces, etc. Diante da grande variedade de sua utilização, as tarefas mais comuns são: classificação, agrupamento, predição (Simon, 2001). Algumas aplicações em Astrofísica foram mencionadas no Capítulo 1.

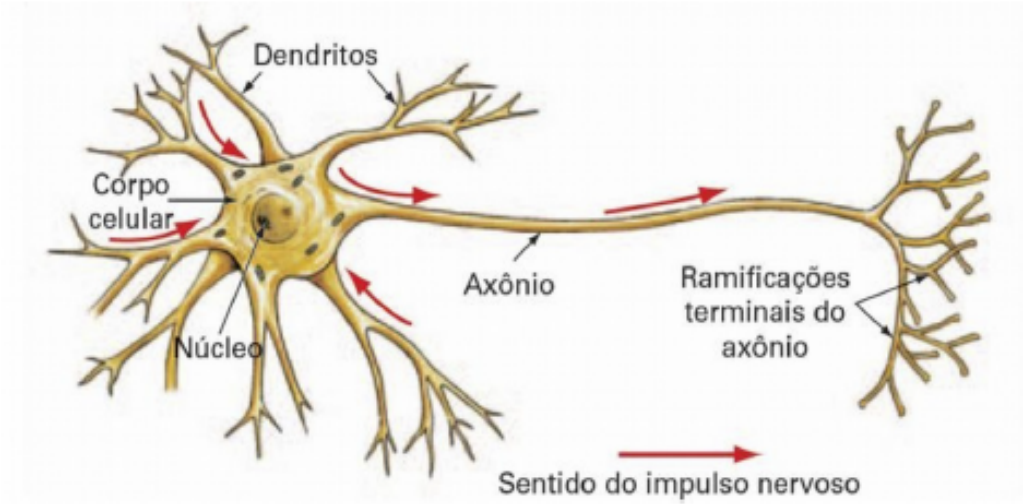
2.1 Tópicos básicos para um projeto de uma RNA

Nessa seção será feita a descrição de tópicos básicos, para a elaboração de um projeto de uma rede neural artificial, com o detalhamento sobre os tipos de topologias e treinamento descritos nas próximas seções.

Os estudos sobre redes neurais artificiais (Kovács, 2002) foram iniciados em 1943 com os trabalhos do médico Warren McCulloch, e do matemático Walter Pitts, que publicaram uma analogia entre as células nervosas e o processo eletrônico, com o desenvolvimento do "neurônio booleano de McCulloc", um dispositivo binário onde as entradas tinham ganhos arbitrários, e as suas somas ponderadas resultavam em uma saída de pulso (excitatória) ou não-pulso (inibitória).

Segundo Favan (2015) em uma rede neural, os neurônios são divididos em três seções: o corpo da célula, os dentritos e o axônio, conforme é apresentado na Figura 2.1. Os dentritos têm a função de receber sinais de outros neurônios, o corpo da célula tem por função de processar esses sinais, que passam pelo axônio, responsável por transmitir os sinais para outros neurônios seguintes. O ponto de contato entre a terminação axônica de um neurônio, e o dentrito de outro neurônio é chamado de sinapse.

Figura 2.1: Modelo da célula neural biológica. Fonte: Adaptado de Favan (2015).



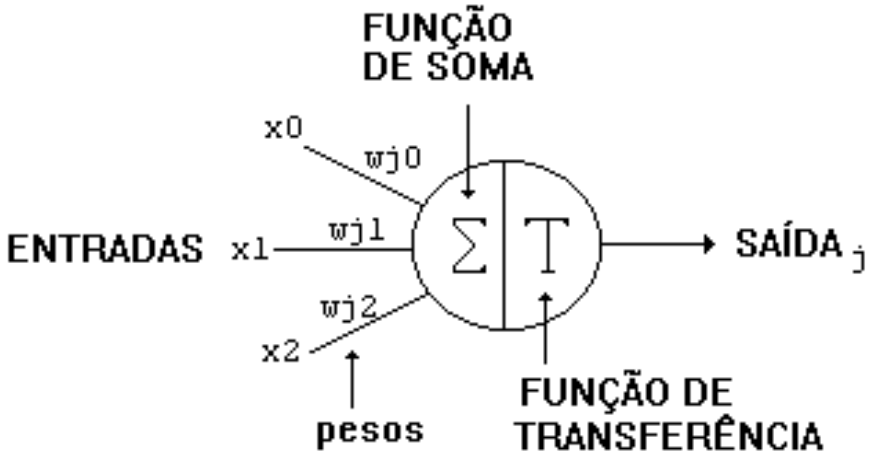
O texto dos parágrafos a seguir dessa seção, quando não mencionada a autoria, foram adptados de Maren et al. (1990). O modelo de neurônio artificial primeiramente proposto por McCulloch e Pitts (1943), é o mais simples, e atende às principais característica do neurônio biológico, sendo utilizado em diversas arquiteturas de redes neurais artificiais. Baseado nesta proposta, as RNAs foram projetadas levando em conta as seguintes hipóteses:

1. O processamento da informação acontece em muitos elementos simples chamados neurônios;
2. os sinais são transmitidos entre os neurônios por conexões de ligações;
3. cada conexão de ligação tem um peso associado, o qual, em uma rede neural típica, multiplica o sinal transmitido;
4. é aplicado em cada neurônio uma função de ativação (usualmente não linear) para sua rede de entrada (soma de pesos dos sinais de entrada) para determinar seu sinal de saída.

Uma RNA é caracterizada por: (1) seu padrão de conexões entre os neurônios (sua arquitetura), (2) seu método de determinar os pesos sobre as conexões (chama-se treinamento, aprendizado ou algoritmo), e (3) sua função de ativação.

Uma RNA consiste de um grande número de elementos de processamento simples de neurônios, unidades, células ou nós. Cada neurônio é ligado a outro neurônio por meio de ligações de comunicação direta, cada um com um peso associado.

Figura 2.2: Neurônio Artificial. Fonte: Adaptado de Maren et al. (1990).



Os pesos representam as informações que são usadas pela rede para solucionar um problema (Simon, 2001). A Figura 2.2 apresenta os componentes de um neurônio artificial, os quais são descritos a seguir:

1. **Entradas e Saídas**

   As entradas ao nó são representadas por um vetor de entrada (Equação 2.1).

$$\vec{x}, Xi(i = 1...N) \tag{2.1}$$

O nó manipula as entradas ou atividade, para a saída Y_j, que pode então ser entrada para outros nós. Outros conceitos são importantes sobre as entradas e saídas da rede:

(a) Dimensão da entrada/saída: quantidade de neurônios de entrada/saída;

(b) codificação: dependente do problema;

(c) normalização;

(d) pré-processamento e pós-processamento.

2. O Fator Peso

É um dos atributos responsáveis pela saída de um neurônio. O fator peso, W_{ij} para a i-ésima entrada, X_i, corresponde ao j-ésimo neurônio. Cada valor de entrada é multiplicado pelo seu correspondente peso, que define o estado do neurônio acessado. A Equação 2.2 define a entrada ao neurônio Y_j:

$$Y_j = X_1 W_1 j + X_1 W_2 j + ... X_n W_n j \tag{2.2}$$

É importante mencionar que os pesos podem ter efeito inibidor ou estimulador. Ajustando W_{ij} tal que $X_i \cdot W_{ij}$ seja positivo (e preferencialmente grande), a tendência é de estimular o neurônio. Se $X_i \cdot W_{ij}$ é negativo, o neurônio é inibido. Finalmente, se $X_i \cdot W_{ij}$ é muito pequeno em magnitude relativa a outros sinais, o sinal de entrada X_i terá pouco ou nenhum efeito.

Tipicamente, os pesos iniciais da rede são gerados aleatoriamente para um dado intervalo.

3. O Limiar Interno

O próximo fator importante para determinar a saída de um neurônio é o seu Limiar Interno. O Componente T_j é o responsável pelo controle da ativação de um dado neurônio, e é denominado Limiar Interno. O Neurônio calcula a soma de todos $W_{ij} \cdot X_i$, e subtrai o valor do seu limiar interno, calculando assim a ativação total conforme é definido pela Equação 2.3:

$$Y_j = \sum_{i=1}^{n} (W_i j X_i) - T_j \tag{2.3}$$

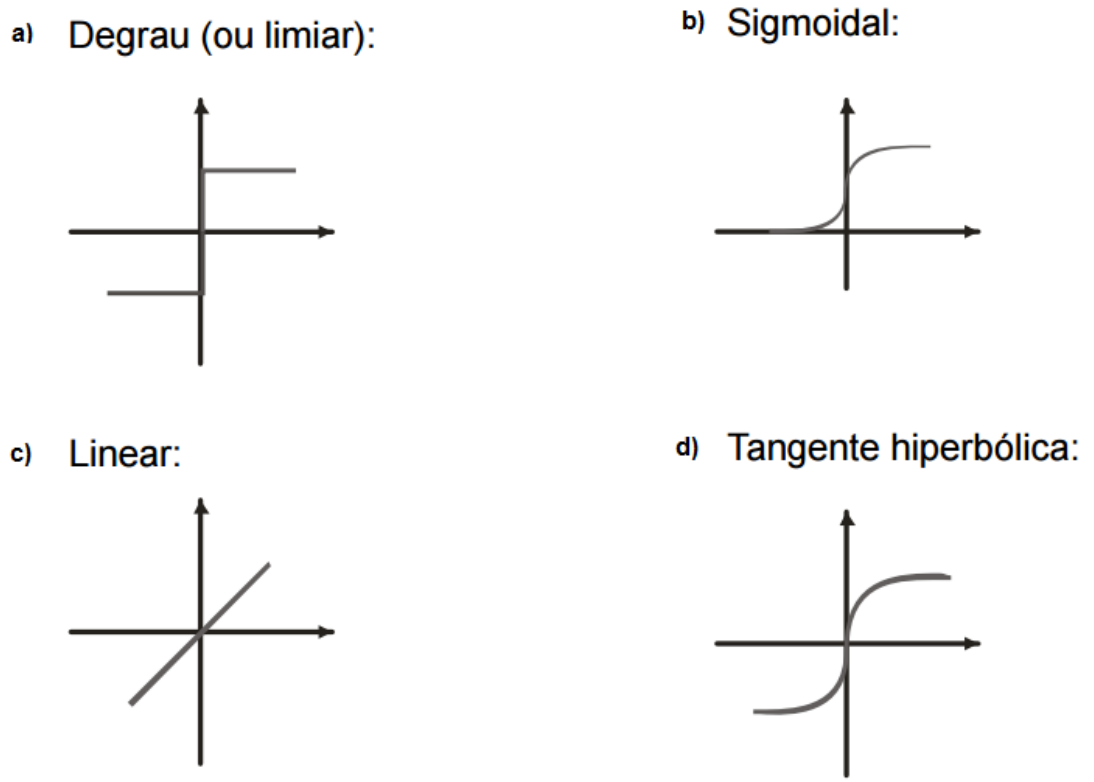
Uma aspecto importante, é que nem todos os neurônios têm um limiar interno. Se nenhum Limiar Interno é especificado, assume-se T_j igual a zero.

4. A Função de Transferência

O fator final, o qual é o responsável pela saída do neurônio é a função de transferência ou ativação. Uma das características mais importantes da função de ativação de um neurônio é que esta deve ser não linear para conseguir aprender mapeamentos não linearmente separáveis.

A Figura 2.3 apresenta alguns tipos mais comuns para funções de transferência. A função limiar assume valores 0 ou 1 para a resposta do neurônio dado um valor de entrada. As funções *sigmóide* apresentam uma curva bem suave e adequada para representar comportamento linear e não-linear.

Figura 2.3: Tipos mais comuns para funções de ativação. Fonte: Adaptado de Maren et al. (1990).

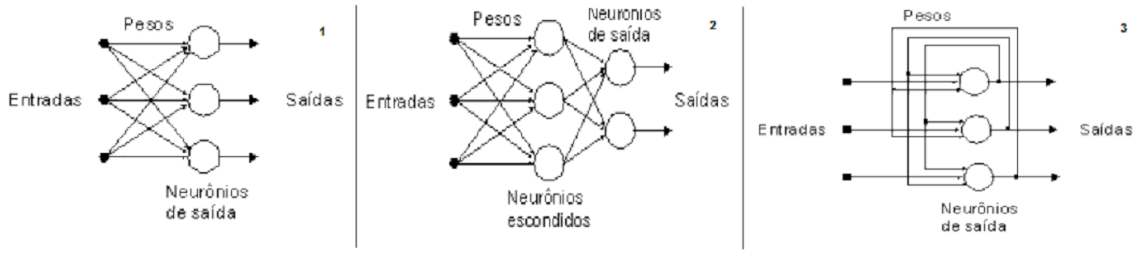


2.2 A topologia de redes neurais

A topologia da RNA refere-se à estrutura da interligação dos neurônios. Essas topologias, ou arquiteturas, se formam organizando os neurônios em camadas, ligando-os à outras camadas. Simon (2001) e Silva et al. (2010) dividem a arquitetura de uma RNA em três tipos: (1) redes *feedforward* de camada simples; (2) redes *feedforward* de camada múltiplas; (3) redes recorrentes.

A Figura 2.4 apresenta essas três topologias citadas. O Tipo 1 é composto de n entradas e m saídas, onde a saída é representada pela própria e única camada de neurônios. O Tipo 2 apresenta uma camada intermediária ou oculta. O Tipo 3 apresenta uma topologia de rede baseada em realimentação, ou seja, as respostas das saídas são entradas para outros neurônios.

Figura 2.4: Exemplos de topologias para redes neurais artificiais: Tipo 1 (esquerda), Tipo 2 (centro), Tipo 3 (direita). Para a definição dos tipos, ver o texto. Fonte: Adaptado de Maren et al. (1990).



2.3 Tipos de treinamento

O texto dessa seção foi adaptado de Silva et al. (2010). Os tipos de treinamento mais comuns para uma rede são o aprendizado supervisionado e o não supervisionado. Em muitas das redes neurais, o aprendizado é a apresentação de uma sequência de vetores de treinamento, ou padrões, cada um com um vetor de saída associado. Os pesos são ajustados de acordo com o algoritmo de aprendizagem. Esse processo é conhecido como treinamento supervisionado.

No treinamento não supervisionado, a rede deve descobrir por si só os padrões, características, regularidades, correlações, ou categorias nos dados de entrada e codificá-los na saída. As unidades e ligações devem transmitir algum grau de auto-organização.

A redundância neste caso é bastante útil para o aprendizado não supervisionado, sem o qual, seria impossível encontrar algum padrão ou características nos dados, o qual poderia parecer apenas um ruído aleatório. Em geral, no treinamento não

supervisionado, o ajuste dos pesos não são realizados tendo como base a comparação com alguma saída desejada.

2.4 Validação cruzada

A técnica de validação utilizada usada na Dissertação consistiu da *K-fold Cross-Validation*, a qual tem por finalidade avaliar a eficácia da generalização do modelo de classificação para novos exemplos. Em um primeiro momento, os dados são particionados aleatoriamente em k amostras. Deve-se ter pelo menos três conjuntos: o conjunto de exemplos de treinamento, validação e teste. O classificador é treinado e validado k vezes (Alpaydin, 2014).

O conjunto de treinamento é utilizado para a aprendizagem do algoritmo, e a amostra de validação para verificar a generalização do algoritmo, conseguido por meio do ajuste dos parâmetros da rede. A etapa final é a da avaliação da generalização sobre o conjunto de testes (Simon, 2001).

O método ainda permite acompanhar a curva de aprendizado durante o treinamento (Figura 2.5). O número de épocas é a quantidade de iterações. A cada época, partições do conjunto de treinamento e validação é processado pela RNA. O erro é acompanhado para cada época. O gráfico de validação cruzada permite observar o melhor critério de parada da rede, que neste caso é representado pelo ponto de parada antecipada.

Figura 2.5: Regra baseada na validação cruzada. Fonte: Simon (2001).

2.5 Métricas de avaliação

Como métrica para avaliar o classificador foram utilizados, os dados da Matriz de Confusão, os quais foram gerados nas fases de treinamento e da classificação da RNA. Posteriormente foram computados os valores da acurácia e Medida-F (*F-Score*).

2.5.1 Matriz de confusão e medidas de avaliação

Dado um classificador e o conjunto de instâncias, uma matriz 2 x 2 denominada matriz de confusão pode ser construída. A Figura 2.6 apresenta um exemplo de uma matriz de confusão, na qual p e n representam as classes verdadeiras; e P e N são as classes preditas pelo classificador hipotético.

Figura 2.6: Matriz de Confusão.

		Classe Verdadeira	
		p - Positivo	**n** - Negativo
Classe Hipotética	**P** - Positivo	**VP** - Verdadeiro Positivo	**FP** - Falso Positivo
	N - Negativo	**FN** - Falso Negativo	**VN** - Verdadeiro Negativo

Com as informações existentes na matriz de confusão, é possível realizar o cálculo de algumas medidas para a avaliação da performance de um classificador, tais como apresentados nas equações 2.4, 2.5, 2.6 e 2.7, respectivamente:

1. Taxa de Verdadeiros Positivos (número de amostras da classe corretamente reconhecida):

$$TVP = \frac{VP}{p} \tag{2.4}$$

2. Taxa de Verdadeiros Negativos (número de amostras que foram corretamente identificadas, como não pertencentes à classe):

$$TFP = \frac{FP}{n} \tag{2.5}$$

3. Precisão (razão entre o número de verdadeiros positivos, e o total de amostras classificadas como positivas):

$$Precisão = \frac{VP}{VP + FP} \tag{2.6}$$

   Obs.: A precisão penaliza os falsos positivos.

4. Revocação (razão entre o número de verdadeiros positivos, e o número de amostras atribuídas a uma determinada classe):

$$Revocação = \frac{VP}{VP + FN} \tag{2.7}$$

   Obs.: A revocação penaliza os falsos negativos.

O uso de métricas permite, a análise do desempenho de determinado classificador.

2.5.2 Medida-F (*F-Score*)

A Medida-F é calculada por meio de duas medidas obtidas, a partir da matriz de confusão e não penaliza somente classificadores com alta taxa de falsos positivos e negativos, mas ambos conforme descrito por Van Asch (2013). A Medida-F é obtida por meio da Equação 2.8:

$$Medida - F = 2 * \frac{Precisão * Revocação}{Precisão + Revocação} \tag{2.8}$$

A Medida-F tem o objetivo de equilibrar tanto a precisão quanto a revocação. Uma observação sobre a Medida-F, é que a mesma não resolve totalmente a questão do desbalanceamento entre classes, porém tenta amenizar um pouco esse problema. A medida F-score é um indicativo de desempenho global do classificador (Van Asch, 2013).

2.5.3 Medida-F (*F-Score*) aplicada para mais de duas classes

Para aplicar a Medida-F para mais de duas classes, é necessário a adaptação do método básico da Medida-F e são necessários os seguintes passos:

1. Construir a matriz de confusão para as diferentes classes;
2. para cada uma das classes calcular as taxas de *VP*, *VN*, *FP*, *FN* em etapas *C*, onde *C* é o número de classes. Em um primeiro momento, altera-se o problema original com uma abordagem em um novo problema com duas classes;
3. para cada classe calcula-se as medidas de precisão e revocação;
4. a Medida-F pode ser obtida da seguintes formas:

   Micro-Média da Média−F = Média das Precisões e das Revocações;

   Macro-Média da Média−F = Precisões e Revocações das Médias dos *VPs*, *FPs* e *FNs*.

Capítulo 3

O Conjunto de dados

O presente capítulo aborda o conjunto de dados utilizado na Dissertação, apresentando o catálogo de aneladas compilado tendo como base outros catálogos, assim como o procedimento adotado para obtermos as imagens de suas respectivas galáxias.

3.1 Os catálogos FAOA e de Moiseev et al. (2011)

Os dois catálogos mais completos que contém galáxias aneladas são os de Faúndez-Abans & Oliveira-Abans (1998, daqui por diante de Catálogo de FAOA) e o Moiseev et al. (2011), contendo 489 e 275 galáxias, respectivamente. O Catálogo de AM87 classifica as galáxias aneladas em três categorias (ver Introdução). Optou-se por utilizar os catálogos de FAOA e o de Moiseev et al. (2011) pois ambos são os mais atualizados do tema.

O Catálogo da morfologia de galáxias peculiares aneladas, de FAOA foi elaborado tendo como base alguns dos objetos dos catálogos de AM87, Few e Madore (1986), Whitmore et al. (1990), FAHR e Faúndez-Abans et al. (1994). As galáxias que foram selecionadas desses catálogos foram inspecionadas tendo como base imagens, nos filtros J e R do levantamento SRC do ESO, com a classificação final em cinco categorias de acordo com a característica geral do anel.

As cinco categorias são: *Polar*, *Hoag*, *Elliptical*, *Irregular*, *Centrally Smooth*. Faúndez-Abans & de Oliveira-Abans (1998) chegaram a essas categorias, tendo como base uma análise prévia, na qual foram considerados distintos tipos de objetos, chegando a um total de 29 tipos de galáxias aneladas. Posteriormente os autores descartaram as galáxias aneladas normais, assim como, agruparam as galáxias com diferentes aspectos morfológicos em uma mesma categoria (ver Introdução). Por sua vez, Moiseev et al. (2011) tendo como base os dados do projeto *Galaxy Zoo* e de outros trabalhos na literatura, compilaram uma lista com 275 galáxias candidatas

aneladas do tipo polar. Os autores também classificaram as galáxias, de acordo com a certeza dos objetos serem aneladas do tipo polares, de acordo com uma classificação de qualidade segundo as suas análises, ou seja, 70 (melhores candidatas), 115 (boas candidatas), 53 (polares, a maioria galáxias com forte presença do *warp* no disco e *mergers*) e 37 galáxias que eles supõem serem polares tendo como base a forte inclinação em relação à linha de visada.

Na Seção 3.4 serão apresentadas figuras, com o exemplo de algumas imagens de galáxias aneladas usadas na Dissertação.

3.2 A compilação do Catálogo *RING−Id* de galáxias aneladas

O Catálogo de FAOA contém 489 galáxias aneladas das cinco categorias principais mencionadas anteriormente, e uma lista de objetos com a flag "?", que significa dúvidas em sua classificação, outro aspecto é a dificuldade de identificar galáxias aneladas irregulares. Descartando ambas as categorias de objetos, restam 307 galáxias. Foi utilizada a versão do Catálogo de FAOA disponível no CDS[1], e por meio de um software escrito em *IDL* realizamos a leitura do mesmo.

Foi realizado procedimento semelhante para a leitura do Catálogo de Moiseev et al. (2011) também disponível[2] no CDS. Optou-se por usar apenas as galáxias classificadas pelos autores com a *flag* ""Best".

Como o Catálogo de Moiseev et al. (2011) de galáxias polares, contém em sua maioria objetos para o Hemisfério Norte não existe superposição de objetos com o Catálogo de FAOA. Em todo o caso, realizamos um cruzamento entre as coordenadas equatorias de ambos os catálogos com um alto valor de raio de procura (1 minuto de arco) e não foram encontrados objetos em comum.

A Tabela 3.1 apresenta o número de galáxias aneladas para cada categoria, tendo como base ambos os catálogos. Nota-se o grande número, 147 e 209, de galáxias classificadas nas categorias Polar (P) e Elíptica (E), respectivamente. Nota-se também, um número reduzido de galáxias existentes para as categorias CS e HL, respectivamente. A quarta coluna, corresponde às galáxias efetivamente selecionadas e o procedimento será discutido na próxima seção. Outro aspecto, é que a única categoria, em que os catálogos FAOA e de Moiseev et al. (2011), tem em comum é a de aneladas polares.

[1] J/A+AS/129/357/table2
[2] J/MNRAS/418/244/table1

Tabela 3.1: Catálogo compilado para o uso na Dissertação.

Autor	Categoria	Quantidade objetos	Selecionadas (RNA)
FAOA	CS	9	7
FAOA	E	209	75
FAOA	HL	12	10
FAOA	P	77	32
Moiseev et al. (2011)	P	70	70

3.3 A obtenção das imagens das galáxias aneladas

Para obtermos as imagens das galáxias para ambos os catálogos (Tabela 3.1) utilizamos o software *Aladin*, que é um Atlas Interativo do Céu desenvolvido e hospedado no Observatório de Estrasburgo na França, e que disponibiliza acesso à milhares de catálogos e possui uma base com imagens de vários levantamento astronômicos.

O melhor comprimento de onda para analisarmos os detalhes das galáxias aneladas corresponde ao do visível, o levantamento mais atual nesse comprimento de onda em grande escala, é o *SDSS* disponível apenas para o Hemisfério Norte.

Dessa forma usamos o *DSS* (*Digital Sky Survey*), que realizou observações no visível para praticamente todo o céu. Utilizamos um script escrito em um outro trabalho do grupo (Cerqueira, 2016) e foram obtidas de forma automática as imagens das 377 galáxias aneladas da nossa amostra (Tabela 3.1), as coordenadas foram obtidas dos catálogos de FAOA e de Moiseev et al. (2011).

As imagens do *Aladin* vem com marca d'água e a marcação da orientação (N/S) nas imagens, assim como a resolução, utilizamos o software *XnView* para retirar essas informações das imagens. Um procedimento similar foi utilizado por Cerqueira (2016). Em uma primeira etapa usamos como tamanho da imagem, dois minutos de arco, algumas imagens possuem tamanho menor do que esse, ou ainda menor do que 1 minuto de arco. Entretanto, para não correr-se o risco de termos imagens cortadas, preferimos usar a priori, imagens quadradas com dois minutos de arco. Identificamos que para muitas galáxias, o centro da imagem não correspondia ao centro da galáxias, estando muitas delas cortadas na imagem.

Por essa razão, verificou-se uma a uma o seu centro, usando o *Aladin* com a sua ferramenta na qual posicionamos o cursor e ele retorna a informação de *RA* e *DEC*. As imagens estão no padrão RGB e representam a composição de três filtros astronômicos no visível (UBV) preferimos trabalhar com imagens em falsa-cor do que monocromáticas para termos maior informação sobre as características do objeto. Quando vemos uma imagem em falsa cor de uma galáxia podemos ver características como por exemplo, a idade média da população estelar, regiões de formação estelar, entre outros aspectos.

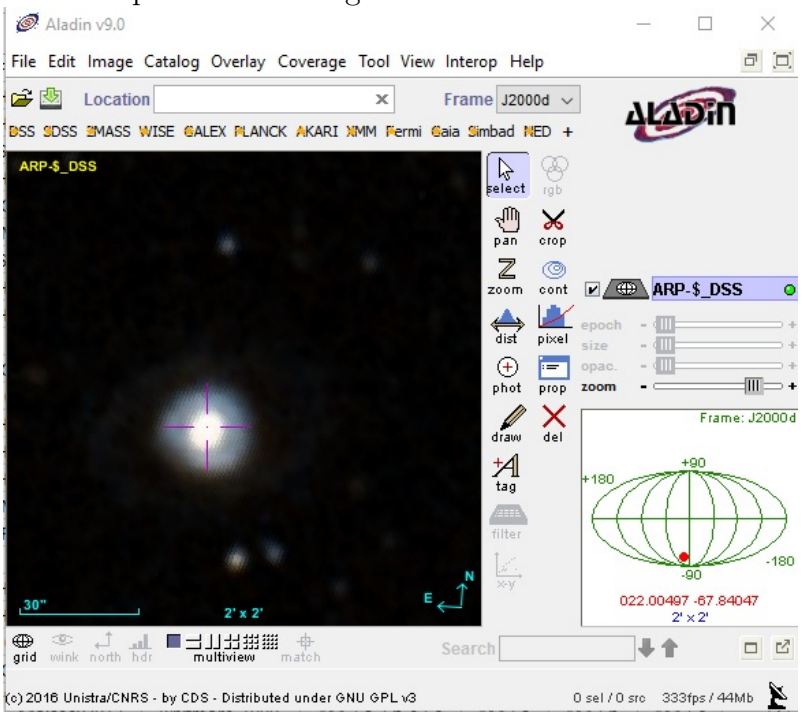
Em uma etapa posterior, foi escrito um outro código que faz a leitura automática das imagens, que estão em uma matriz de 300 x 300 e transforma em um arquivo

ASCII, com os valores de frequência em que cada ponto de interesse é detectado e descrito com relação a distância de alguns outros pontos gerados pelo algoritmo *K-means*. Detalhes do funcionamento do algoritmo, que detecta e descreve cada ponto de interesse em imagens e da elaboração dos histogramas de frequências são apresentados no Capítulo 4.

A Figura 3.1 apresenta uma tela do software *Aladin* na obtenção de galáxias do estudo. A figura apresenta uma imagem de galáxia com 2 minutos de arco obtida pelo *Aladin* da galáxia *AM 0126-680*. Nessa imagem também é possível ver o tamanho, a escala e a orientação da mesma. No canto direito é apresentada a localização de suas coordenadas, na esfera celeste, em coordenadas equatoriais.

Podem também ser observadas estrelas (objetos brilhantes com formato pontual) que estão entre nós e a galáxia, muito provavelmente estrelas brilhantes. Nota-se uma pequena diferença entre o retículo do *Aladin* com o centro da imagem e o centro da galáxia, o que deve-se ao fato de que obtemos as coordenadas do *NED*. Essa pequena diferença, não interfere em nosso estudo, nem tampouco as estrelas de campo, pois fazemos um corte a partir do centro, de forma com que a imagem final contenha apenas o objeto de estudo, salvo algumas exceções de alguns objetos muito próximo da galáxia. Cabe ressaltar, que esse foi um dos critérios, levados em consideração para o descarte de alguns objetos utilizados em nossa amostra.

Figura 3.1: Exemplo de uma imagem obtida com o uso do software *Aladin*.



Dessa forma, foi realizado o download das imagens das galáxias existentes nos catálogos FAOA e Moiseev et al. (2011), que contém 307 e 70 objetos respectivamente.

3.4 Seleção das imagens a serem usadas

Conforme mencionado na seção anterior, há uma quantidade considerável de imagens nas quais existe uma quantidade considerável de estrelas, as quais irão afetar a nossa amostra de treinamento, fazendo com que as suas características também estejam presentes e não apenas as das galáxias aneladas. Em uma etapa futura, pretendemos usar o método que está sendo usado em outro trabalho do grupo (Cerqueira et al. 2018, em preparação) no qual de forma automática identificamos as estrelas e as retiramos de modo artificial das imagens.

No presente trabalho foi realizado um procedimento, manual, ou seja, foi feito um recorte na forma de um quadrado, a partir do centro dos objetos com o objetivo de retirar as estrelas da imagem de dois minutos de arco, quando a estrela estava muito próximo no plano xy da galáxia, simplesmente descartamos a imagem. Manualmente, verificou-se cada centro do objeto utilizando ferramenta *Aladin* e uso do software *ImageJ* para o recorte.

Após esse procedimento de seleção dado o baixo número de galáxias (Tabela 3.1) das categorias *CS* e *HL*. Tendo em vista, o reduzido número de objetos nas outras categorias, optou-se pelo uso de apenas, as categorias elípticas e polares. A Tabela 3.2 apresenta um resumo da quantidade de objetos, selecionados para a RNA em cada amostra das categorias Polares e Elípticas.

Tabela 3.2: Objetos selecionados e que serão usados na RNA.

Objeto	Selecionadas (RNA)
Elíptica (E)	75
Polar (p)	102

Tendo como base as imagens selecionadas, foram compiladas amostras de treinamento e classificação, com um total de 124 e 53 galáxias, respectivamente. Nas figuras 3.2 e 3.3 são apresentados alguns objetos representativos da categoria Elíptica e Polar, após o processo de seleção.

Figura 3.2: Objetos representativos da Categoria Elíptica.

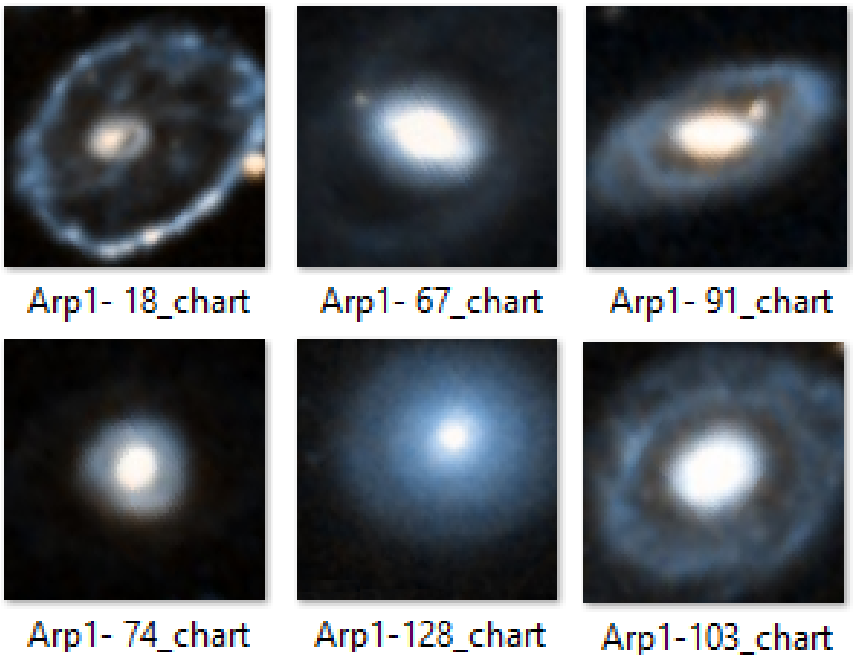
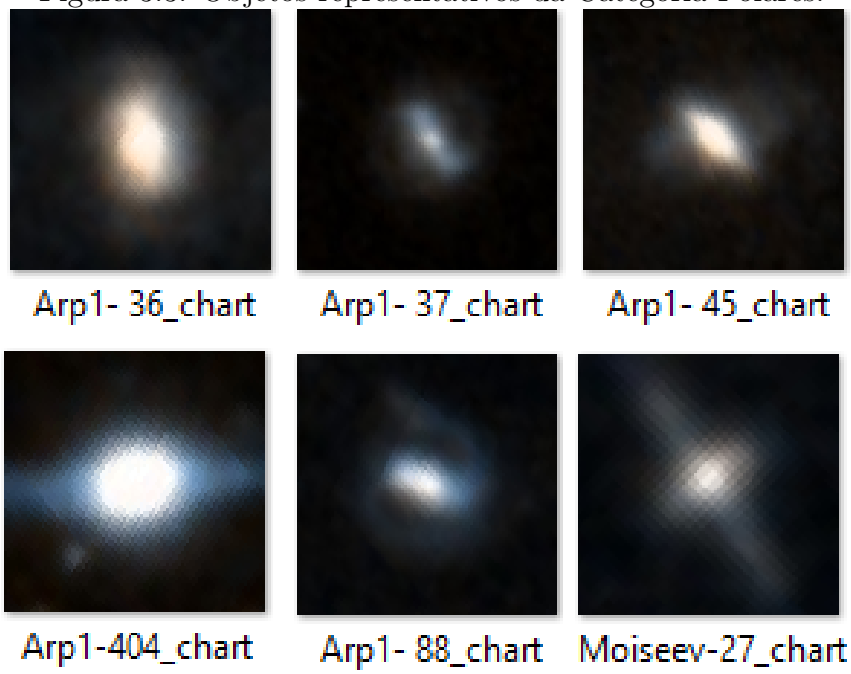


Figura 3.3: Objetos representativos da Categoria Polares.



Com as imagens selecionadas foi escrito um outro código, com o uso da biblioteca *LIRe* para elaborar as entradas para a nossa RNA. O código faz a leitura de forma automática de imagens de galáxias e extrai descritores utilizando o algoritmo *SURF*, descrito no Capítulo 4.

Capítulo 4

Metodologia

No capítulo é descrita a metodologia empregada na Dissertação, que consiste na definição dos tipos de galáxias aneladas do estudo, a obtenção de imagens astronômicas de galáxias aneladas peculiares, o tratamento das imagens selecionadas, a identificação e descrição das características dessas imagens, por meio da utilização da biblioteca *LIRe*, a elaboração de um software ($RING-Id$) de RNA, a execução das etapas de treinamento, validação e testes da rede.

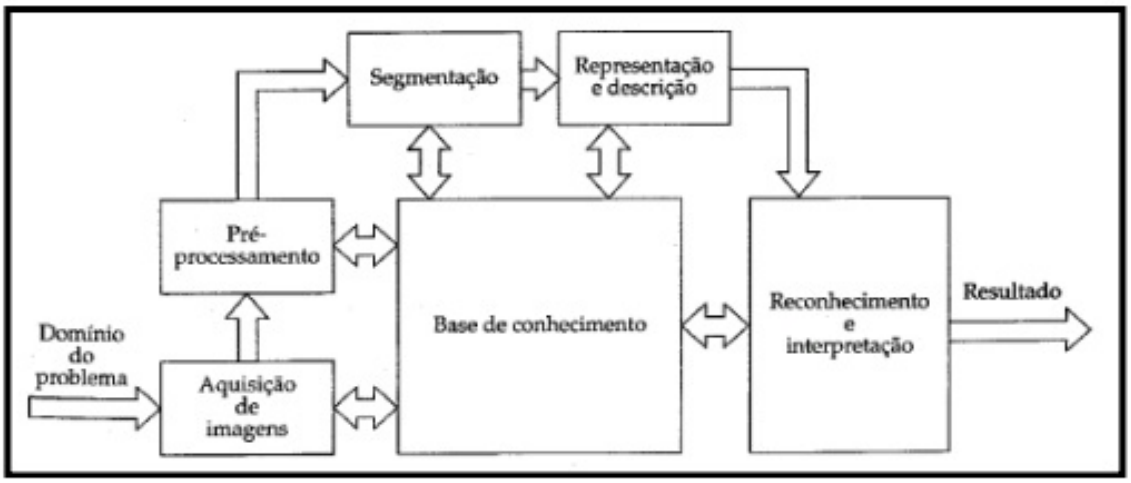
Realizou-se também um estudo sobre a viabilidade do tratamento das imagens obtidas com o *Aladin*. Em uma primeira etapa preferiu-se usar as imagens originais, e verificar os resultados, para em uma etapa futura implementar, caso necessário o processamento das imagens. Notou-se que durante a fase de treinamento utilizando as imagens originais, a rede demorava para convergir. Dessa forma, optou-se por realizar experimentos utilizando técnicas de extração de atributos locais em imagens, visando o baixo custo computacional com bom desempenho.

4.1 Extração de atributos em imagens

O processamento de imagens surge com o objetivo de extrair informações de interesse, tornando-se assim uma ferramenta essencial no mundo moderno, e que vem sendo amplamente utilizado em diversos setores (Gonzalez e Woods, 2012). Conforme foi visto na Introdução, na Astronomia de forma particular, muitos estudos estão sendo realizados na área de reconhecimento de padrão, o qual requer informações sobre imagens.

Para Gonzalez e Woods (2000) existem várias etapas no processamento de imagens, que vão desde a sua aquisição até a sua interpretação. Os passos essenciais no processamento de imagens são descritos no fluxograma apresentado na Figura 4.1.

Figura 4.1: Etapas principais em processamento de imagens digitais. Fonte: Gonzalez e Woods (2000).



Dentre as diversas fases do processamento digital de imagens, a etapa do pré-processamento é uma peça fundamental, pois na fase de aquisição, ruídos podem estar presentes causando certas distorções nas imagens coletadas.

A fase de como descrever a imagem em um conjunto de atributos, depende muito do domínio do problema, mas geralmente procura-se por características que sejam invariantes a possíveis transformações que a imagem possa sofrer, como translação, mudança no tamanho e rotação (Dougherty, 2009).

A extração de características em imagens utiliza duas perspectivas, as locais e as globais. Os atributos globais em imagens são a cor, textura e forma. Os descritores globais não são muitos bons, quando uma dada imagem altera a sua perspectiva, como por exemplo, ser rotacionada ou transladada (Lisin et al., 2005). Diferente dos atributos globais, os locais são descritos com um padrão existente na imagem associada, por exemplo, a alteração de perspectiva, podendo ser um ponto de interesse (Lisin et al., 2005).

No presente trabalho buscamos identificar as características em imagens de diferentes tipos de galáxias aneladas peculiares, com a utilização de alguns descritores locais utilizando o Algoritmo *Speeded Up Robust Features* (SURF) e o emprego da técnica para a descrição de imagens baseados em *Bag of Features* (BoF).

4.2 A *LIRe*

A *LIRe* (*Lucene Image Retrieval*) é uma biblioteca Java distribuída sob a licença GNU GPL, que fornece uma série de recursos para a extração de características globais e locais de um arquivo de imagens, e possui recursos de indexação, para encontrar similaridade em imagens. De acordo com Penatti et al. (2012) essa opção pode ser estudada por meio de alguns dos seus descritores, ou seja, o quanto um dado descritor é parecido com outro.

Ela é baseada em código aberto, múltiplas plataformas e de fácil adaptação, além de ser muito flexível, e de fácil implementação, sendo dessa forma muito atrativa aos utilizadores, que se aproximam pela primeira vez do tema sobre recuperação de informação em imagens.

A primeira versão da biblioteca foi disponibilizada em 2004, fazendo parte do projeto Caliph & Emir, que visava fornecer recursos para a recuperação de imagem baseada em conteúdo e métodos para pesquisa de índices (Gonçalves, 2016). Nesse processo de recuperação é calculada a distância dos descritores gerados e indexados. A métrica da distância adotada, a exemplo da distância euclidiana, mostra o quanto um dado vetor de característica é similar à um outro vetor (Gonçalves, 2016).

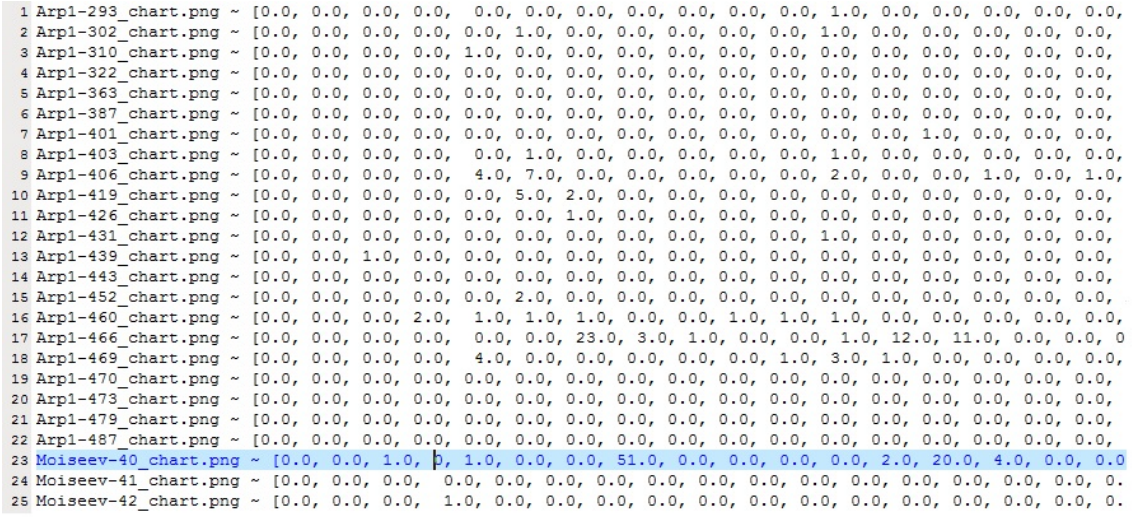
A LIRe fornece classes e interface importantes, como é o caso da classe *ParallelIndexer*, que implementa métodos importantes para a indexação de imagens. Um exemplo é o próprio construtor da classe, que a partir de um diretório contendo imagens, nos permite modificar parâmetros, como por exemplo, adicionar extratores para criar um arquivo base das características encontradas em cada imagem.

O algoritmo que realiza o processo de indexação, adiciona alguns descritores. No algoritmo é necessário criar o arquivo de imagens e informar o seu caminho no disco rígido de um computador. A classe *ParallelIndexer* recebe os parâmetros, como quantidade de imagens e o tamanho do vocabulário, assim como, o extrator utilizado representado pelo Algoritmo *SURF*, que tem como parâmetros: *hessianThreshold*, *nOctaves*, *nOctaveLayers*, *extended* e *upright* descritos nas próximas seções.

Em uma etapa posterior é feita a busca de similaridade, entre o conjunto total de imagens, e a imagem de interesse ou de entrada. Dada uma métrica de distância, a exemplo das distâncias de Manhattan, Euclidiana, dentre outras. As imagens são listadas apresentando os atributos *score*, e o nome do objeto. Quanto mais próximo for de zero, o *score*, mais similaridade as imagens terão entre si.

A Figura 4.2 apresenta a estrutura do arquivo onde foram gerados os histogramas para as imagens. A imagem foi cortada para melhor visualização. Em cada linha do arquivo, existe o nome da galáxia, com os 256 valores, que representam os pontos de interesse detectados e descritos, o qual é associado à um determinado ponto gerado pelo o Algoritmo *k-means*. Mais detalhes dos algoritmos mencionados são discutidos nas próximas seções.

Figura 4.2: Arquivo que contém todos os histogramas.



Dentre os recursos disponíveis na *LIRe*, podemos destacar:

1. *ColorLayout* e Histograma Edge MPEG-7;
2. Descritores de cor: *CEDD and FCTH* (contribuição de Savvas Chatzichristofis);
3. histogramas de Cores (*HSV* e *RGB*), Tamura & Gabor;
4. histogramas de Palavras Visuais por meio do *SIFT* ou *SURF*;
5. indexação e Pesquisa.

No presente trabalho foi utilizado, o Algoritmo *SURF* implementado na biblioteca LIRe, tendo sido base para a detecção e descrição dos pontos de interesse, em um banco de imagens astronômicas.

4.3 *Speeded Up Robust Features (SURF)*

Proposto por Bay et al. (2008) o Algoritmo *SURF* é um sólido detector de pontos de interesse em imagens. O SURF é um algoritmo para detecção e descrição de pontos de interesse (*key points*) e foi baseado em outro algoritmo denominado *Scale Invariant Feature Transform* (*SIFT*), contudo, o *SURF* é mais eficiente na extração de características (Bay et al., 2008). O Algoritmo *SURF* assim como o *SIFT* faz uso de dois princípios fundamentais descritos nas próximas duas subseções. O texto das subseções 4.3.1 e 4.3.2 foi adaptado de Bay et al. (2008).

4.3.1 Detecção de pontos de interesse

Para a fase da detecção dos pontos de interesse em imagens, o Algoritmo *SURF* utiliza de alguns conceitos, os quais tem por objetivo encontrar pontos interessantes que sejam relevantes em uma imagem, e que permita a comparação com outras imagens para encontrar alguma similaridade. A seguir são listados os conceitos importantes para a sua melhor compreensão:

1. Matriz Hessiana e Filtros de caixa

   As matrizes Hessianas são importantes instrumentos para detectar ocorrência de variações em imagens. Formadas pelos conceitos das derivadas parciais nas direções x, y e na direção da diagonal z dado um determinado ponto e pelo cálculo do determinante que nos permite saber a intensidade da variação.

   De acordo com Freitas (2015), a Equação 4.1 mostra para cada posição da matriz a variação dos pixels em uma determinada orientação, obtidos por meio da técnica de convolução.

$$H(P,\sigma) = \begin{bmatrix} L_{xx}(P,\sigma) & L_{xy}(P,\sigma) \\ L_{xy}(P,\sigma) & L_{yy}(P,\sigma) \end{bmatrix} \tag{4.1}$$

   A posição L_{xx} corresponde a variação na horizontal, L_{yy} a variação na vertical e as diagonais formadas pelas variações em L_{xy}. A variável σ corresponde ao valor da escala.

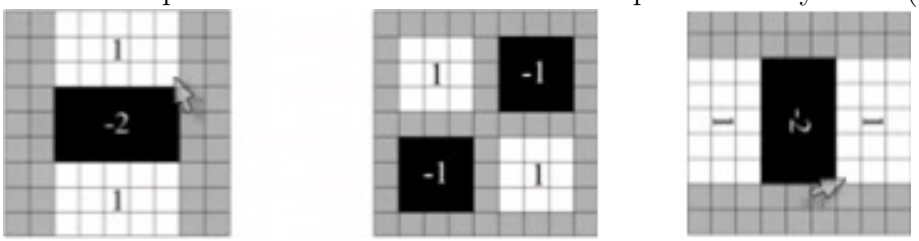
   Uma forma de saber se existe a variação, consiste em calcular os valores correspondentes às posições da Matriz Hessiana, e de encontrar o valor do determinante da matriz. Quando o valor do determinante é relativamente baixo significa que houve pouca variação.

   Para o cálculo do determinante, o *SURF* faz uso da Equação 4.2.

$$det(H_approx) = D_{xx}D_{yy} - (wD_{xy})^2 \tag{4.2}$$

Os valores aproximados das derivadas parciais da equação, em um certo ponto são obtidos por meio da técnica de convolução utilizando filtros com os núcleos conforme apresentado na Figura 4.3.

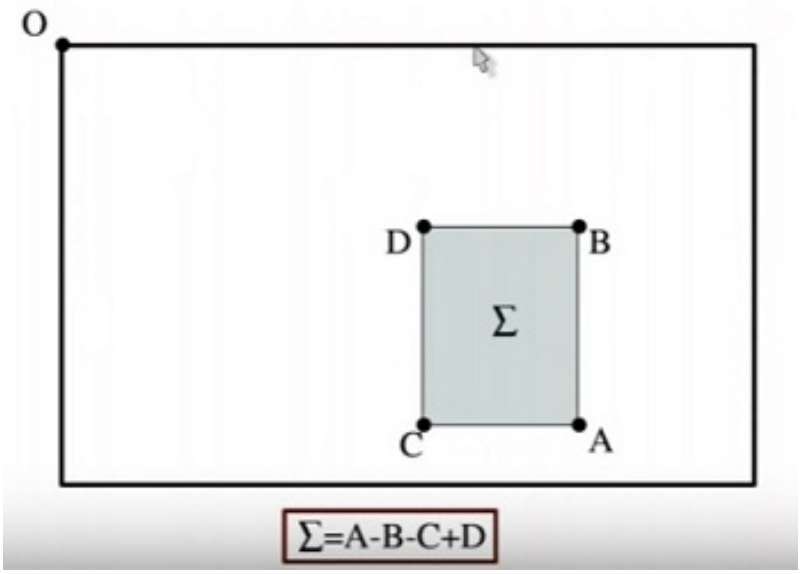
Figura 4.3: Exemplo de filtros de caixa. Fonte: Adaptado de Bay et al. (2008).



As matrizes de filtros conforme apresentado na Figura 4.3 percorrem uma região da imagem pixel a pixel nas orientações x, y e z aplicando os seguintes valores de núcleos: a esquerda, centro e direita respectivamente. O resultado desta operação é dado como a resposta da convolução. Um valor de resposta relativamente alto significa que existe uma maior variação naquela região na imagem original.

Uma observação é que o valor da variável w é uma constante, que é pré-calculada, e serve para compensar as diferentes escalas.

Figura 4.4: Exemplo de imagens integrais. Fonte: Adaptado de Bay et al. (2008).



2. Imagens integrais

O objetivo das imagens integrais é propiciar uma melhor eficiência do Algoritmo *SURF*. A Figura 4.4 apresenta como são calculados os resultados de uma

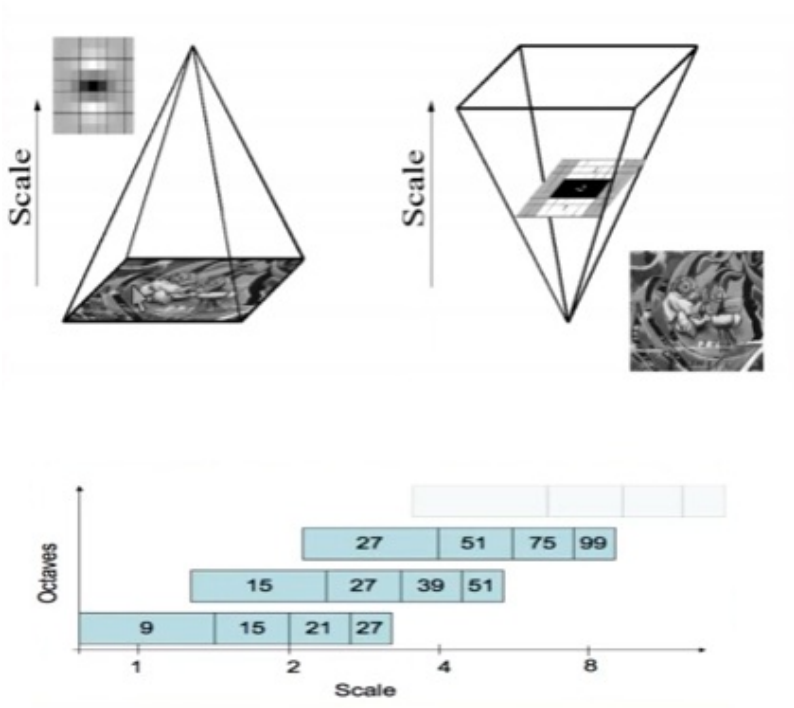
matriz, representada pelo retângulo menor baseado na imagem original, que é representado pelo retângulo maior. Cada elemento da nova matriz, a exemplo do ponto D da figura será o valor do somatória correspondente, a todos os valores de tons de cinza da região que está na horizontal deste ponto e dos valores da vertical.

Com o conceito de imagens integrais, as operações de convolução são mais simples do que o trabalho direto nas imagens originais.

3. Pirâmide de escalas e Oitavas

   As pirâmide de escala é necessária quando há alteração de tamanho de uma imagem. A Figura 4.5 mostra diferentes escalas para uma dada imagem, quando houve uma mudança de tamanho da imagem original. Uma nova escala é obtida por meio do aumento da dimensão dos filtros. A terminologia *Octaves* na figura é um conceito que define escalas intermediárias.

Figura 4.5: Pirâmide de escalas sem reduzir o tamanho das imagens. Fonte: Adaptado de Bay et al. (2008).

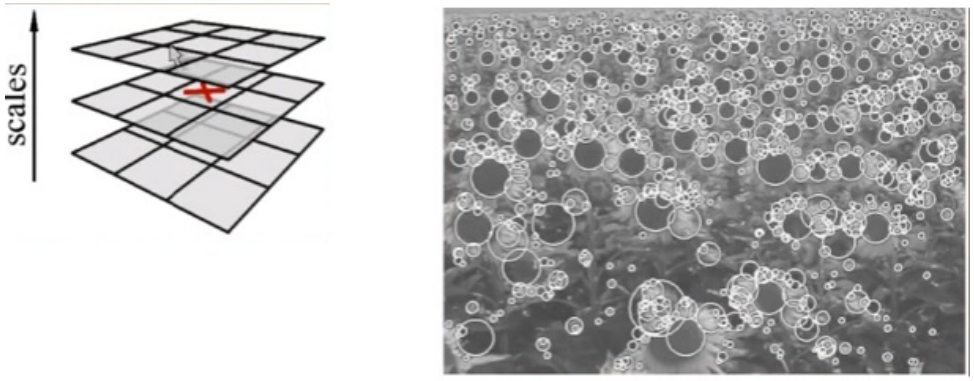


4. Supressão de não máximo em 3D

   Continuando com o objetivo de encontrar pontos de interesse. A supressão de não máximo em 3D, é o método da verificação dos pixels da vizinhança de uma imagem e de escalas imediatamente inferior e superior, ou seja, filtros de tamanho maior e menor aplicado a imagem. Procura-se o ponto de maior valor ou variação, e descarta-se outros pontos de valores menores. O *SURF* informa

a posição em que o ponto de interesse e escala foi encontrado. A Figura 4.6 apresenta o conceito da supressão de não máximo em 3D.

Figura 4.6: Supressão de não máximo em 3D. Fonte: Adaptado de Bay et al. (2008).



4.3.2 Descrição dos pontos de interesse

Nessa fase o Algoritmo *SURF* extrai informações, de cada ponto de interesse detectado na fase anterior e adiciona para este processo, novas ações apresentadas nas figuras 4.7, 4.8 e 4.9 com a finalidade de gerar um vetor de atributos que descreve cada ponto de acordo conforme os seguintes passos:

1. O cálculo de orientação dominante para invariância à rotação;

   Quando uma dada imagem sofre uma rotação, é interessante que o vetor de atributos continue sendo o mesmo. O objetivo é encontrar esse mesmo vetor, em uma outra imagem parecida da imagem original que sofreu rotação. O vetor de atributos é o um descritor do ponto de interesse.

   Para conseguir o objetivo descrito, é aplicada novamente a convolução, tendo como base dois filtros que identificam variações numa determinada orientação. Para cada ponto de interesse, o Algoritmo *SURF* seleciona uma região, cujo tamanho está relacionado com a escala e faz o cálculo da resposta da convolução para os dois filtros conforme exemplo da Figura 4.7.

   Os dois quadrados menores da figura são filtros aplicados em cada ponto de interesse. A parte escura do quadrado forma um conjunto de valores iguais a 1, e a parte clara igual a -1. Os filtros são aplicados na direção x e y. A imagem do círculo mostra, onde é encontrado o maior da variação na imagem, dado um ponto de interesse.

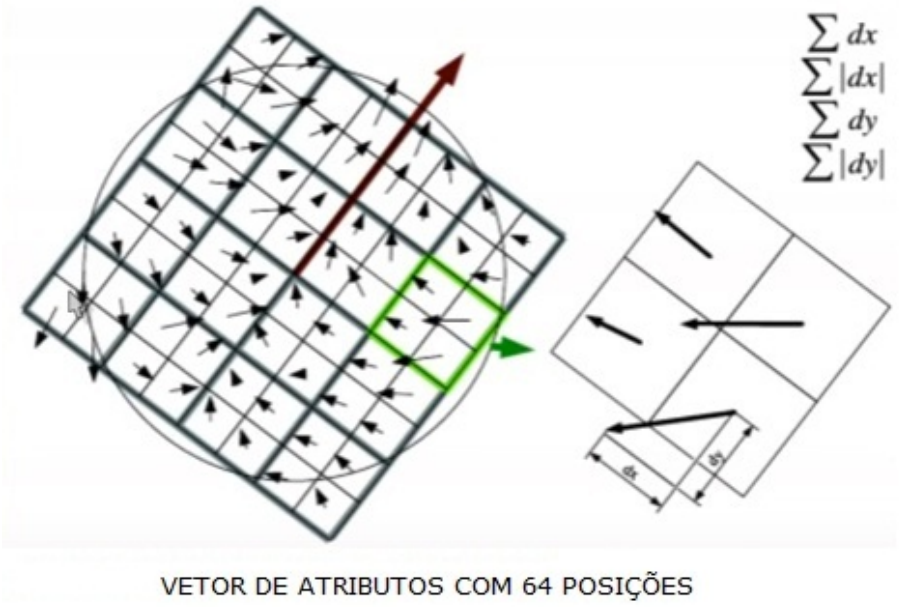
2. Distribuição de respostas de *Haar Wavelets*;

   Após o calculo das direções, o Algoritmo *SURF* realiza um alinhamento de uma área quadrada na esquerda da Figura 4.8 com uma área que está diretamente

Figura 4.7: Cálculo de orientação dominante. Fonte: Adaptado de Bay et al. (2008).

relacionada à escala, e com o ponto de interesse detectado numa determinada direção, e, logo em seguida, divide essa área em quadrados menores, num total de 16. Para cada nova área é calculada a resposta da convolução, utilizando novamente os dois núcleos já mencionados no passo anterior que fornece informação nas orientações x e y.

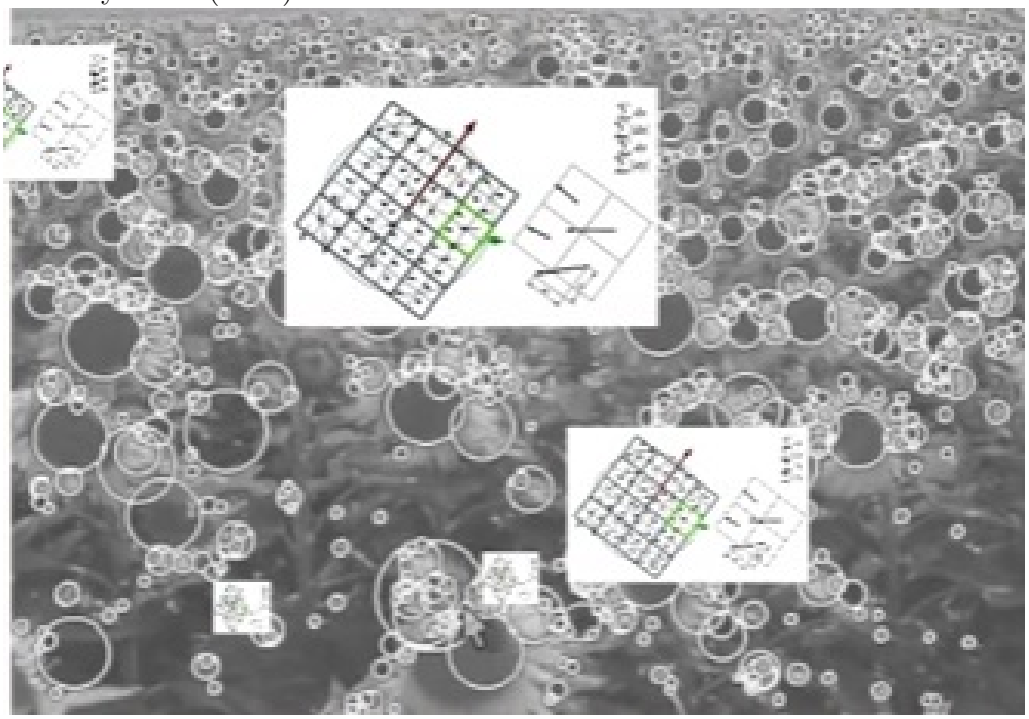
Figura 4.8: Distribuição de respostas de *Haar Wavelets*. Fonte: Adaptado de Bay et al. (2008).



3. Vetor de 64 atributos para cada ponto de interesse.

   Para cada ponto de interesse um descritor de 64 posições é calculado, o qual independe da rotação. Os atributos mostram um padrão na orientação. A Figura 4.9 mostra os vetores que representam cada ponto de interesse da imagem.

Figura 4.9: Descritores com 64 posições para cada ponto de interesse. Fonte: Adaptado de Bay et al. (2008).



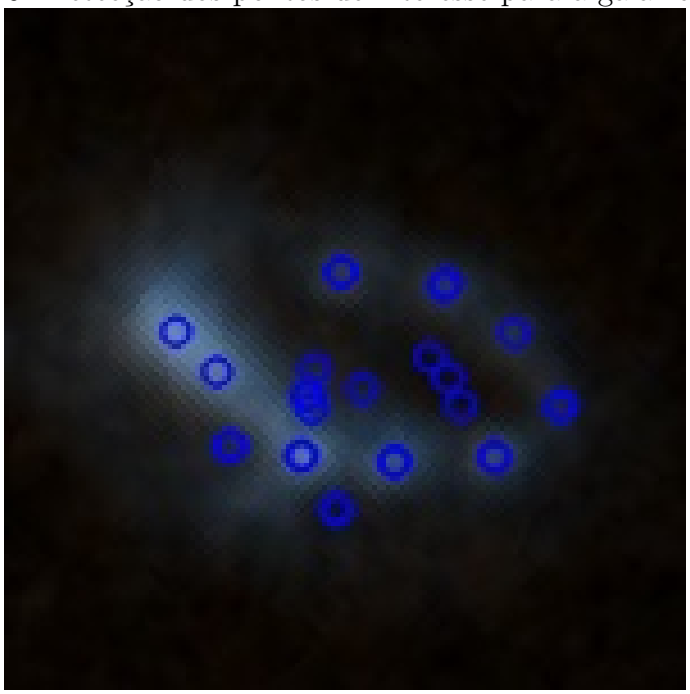
Ao utilizar o *SURF*, alguns atributos são essenciais, a exemplo do parâmetro limiar *hessianThreshold*, que gera influência sobre a quantidade de pontos de interesse retornados, ou seja, quanto menor esse atributo mais pontos de interesse serão apresentados (Laganière, 2014). A Tabela 4.1 apresenta parâmetros do algoritmo *SURF*.

Tabela 4.1: Parâmetros do Algoritmo *SURF*.

Parâmetro	Conceito	Faixa
hessianThreshold	Limiar	0 - 255
nOctaves	Quantidade de oitavas na pirâmide utilizada	1 - 5
nOctaveLayers	Quantidade de camadas em cada oitava	1,2,4,8
extend	Tipo do descritor	128, 64
upringht	Rotação	falso

A Figura 4.10 apresenta um exemplo de informações dos pontos de interesse detectados, com o uso da ferramenta *LIRe* e do Algoritmo *SURF* para uma dada imagem de galáxia.

Figura 4.10: Detecção dos pontos de interesse para a galáxia Arp1-34.



4.4 Bag-of-Features (BoF)

O *Bag-of-Features* também conhecido como *Bag-of-Keypoints* utiliza o conceito da técnica de *Bag-of-words*. A técnica foi utilizada a princípio para a análise de documentos de texto, e posteriormente para aplicação em visão computacional. Para imagem, a regra é similar às utilizadas em texto, quantificando regiões descritas (Bosch et al., 2007).

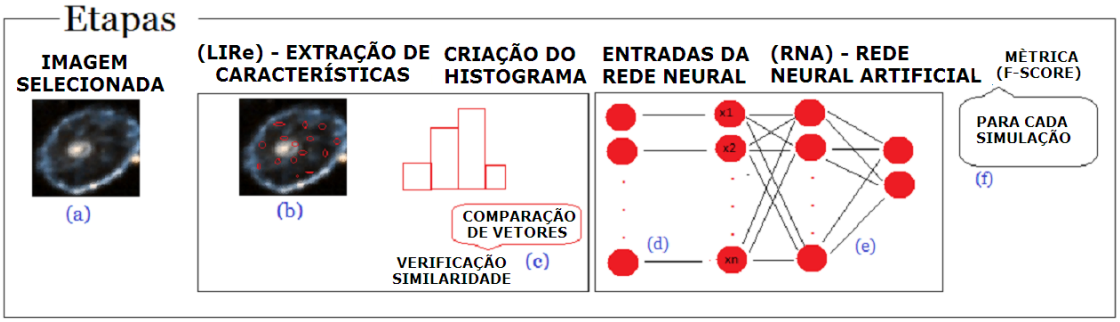
Na Dissertação a técnica *BoF* foi utilizada para criar histogramas de palavras visuais, baseadas nas descrições dos pontos de interesses detectados. Para montar os histogramas foram adotados diferentes tamanhos de dicionários. Para a criação do vocabulário a biblioteca *LIRe* foi utilizada seguindo os procedimentos listados a seguir:

1. Detecção e Descrição dos pontos de interesse nas imagens por meio do Algoritmo *SURF*;
2. construção do vocabulário a partir dos pontos de interesse utilizando o algoritmo de clusterização *k-means*. Onde K é uma constante que define o tamanho do vocabulário utilizado. Mais detalhes sobre o algoritmo *K-means* consultar (Pham et al., 2005);
3. construção do histograma de frequência.

Na *LIRe* os parâmetros *aggregator* do método *ParallelIndexer* permitem indexar e recuperar em formato de um histograma uma dada imagem. Para isso, a classe *ParallelKMeans* é utilizada para identificar, a menor distância entre os centros gerados pelo algoritmo *k-means*.

O passo (c) da Figura 4.11 mostra um exemplo de criação de um histograma de palavras visuais. Dada uma imagem de galáxia (a) é capturado os pontos de interesses dessa imagem (b), os quais são descritos em um vetor de atributos, e a técnica *BoF* é utilizada para a criação de um histograma baseada nas descrições dos pontos de interesse.

Figura 4.11: Histograma de palavras visuais.



Cabe ressaltar, que o histograma de palavras visuais será transformado em um vetor de atributos, que será utilizado como entradas do modelo inferido pelo algoritmo de aprendizado automático conforme passo (d) da Figura 4.11.

4.5 Medidas de similaridade em imagens

Medidas de similaridade são utilizadas para calcular, o quanto uma imagem é semelhante à outra imagem da massa de dados. Essas métricas utilizam a representação vetorial das imagens para determinar a distância entre elas (Sales e Calumby, 2010).

Logo após a fase da elaboração dos histogramas de palavras visuais, são usadas as medidas de distância com o objetivo de encontrar objetos similares. Métricas como a distância euclidiana, podem ser adotadas para o cálculo de similaridade entre imagens conforme é apresentado na Equação 4.3.

$$d(d_1, d_2) = \sqrt{\sum_{i=0}^{n} (w_{1i} - w_{2i})^2} \tag{4.3}$$

Na expressão acima, $d(d_1, d_2)$ é a distância entre duas imagens d_1, d_2, na qual n é o tamanho do vetor que representa cada imagem e w_{1i} e w_{2i} são os valores das palavras visuais de duas imagens hipotéticas, respectivamente na posição i.

4.6 A extração de características de galáxias aneladas

No presente trabalho foi utilizado, o Algoritmo *SURF* implementado na biblioteca *LIRe*, os quais constituíram-se em ferramentas fundamentais para a detecção e a descrição de pontos de interesse em um conjunto de imagens, no presente estudo de imagens astronômicas.

Na manipulação de imagens astronômicas, mesmo com galáxias em um campo pequeno, mesmo sem interação com outras galáxias, existe muitas vezes a presença de estrelas, isso deve-se ao fato, dessas estrelas serem na maioria dos casos, objetos que estão localizados dentro de nossa Galáxia, e estão entre nós, e as galáxias de interesse.

Dessa forma, em uma primeira etapa para algumas imagens obtidas por meio do software *Aladin*, notou-se a existência de uma quantidade significativa de estrelas nas imagens. Como essas estrelas não são objetos do presente estudo, e não possuem obviamente nenhuma ligação com as galáxias (estão no mesmo plano XY, entretanto, estão localizadas à distâncias extremamente díspares, entre si) nas quais iremos fazer o reconhecimento de padrões, optamos por descarar essas imagens.

Como o número de galáxias no presente trabalho não é alto, optou-se por simplesmente recortar as imagens com o intuito de descartar as estrelas das mesmas. Dessa forma, busca-se o centro da galáxia de cada imagem, e é feito um recorte em forma de quadrado com um tamanho de lado variando entre as imagens. Para efetuar tal procedimento utiliza-se o software *ImageJ* e a função *crop* implementada no mesmo. Após os recortes utilizamos a ferramenta *LIRe* para detectar os pontos de interesse, para cada imagem, e obter-se descrição de cada um destes pontos em vetores de 64 posições.

Tabela 4.2: Definição da faixa de parâmetros.

Nome	Sigla	Faixa
hessianThreshold	*l*	0,10,50, 100, 150, 200, 250, 300, 350, 400, 450 e 500
Octaves	*o*	1, 2, 3, 4, 5, e 6
nOctaveLayers	*nO*	2, 3, 4, 5, 6, 7, 8 e 20
extended	*e*	false
upright	*u*	false

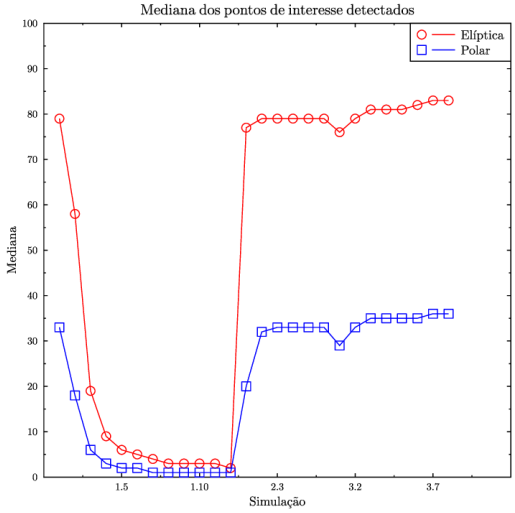
Foi utilizada também uma técnica conhecida como *BoF*. Conforme mencionado anteriormente, a referida técnica cria vários histogramas de palavras visuais, baseadas nas descrições dos pontos de interesses detectados. Para montar os histogramas são definidos diferentes tamanhos de dicionários com dimensão de 32, 64, 128 e 256.

Para identificar se uma dada imagem era de uma galáxia anelada dos tipos abordados no presente estudo (polares e elípticas), foi elaborado um software para o reconhecimento de padrão. Nessa fase, o software realizou uma classificação automática usando como parâmetro, os próprios histogramas que serviram como indutores de uma RNA.

Para detectar os pontos de interesse em imagens de galáxias, foram ajustados alguns parâmetros do Algoritmo *SURF* com o objetivo de encontrar uma boa quantidade com uma quantidade adequada de pontos de interesse. Os parâmetros ajustados são descritos a seguir:

1. *hessianThreshold*: limiar da detecção Hessiana. Apenas os pontos com resposta Hessiana maior do que este limiar são considerados. Quanto maior for este limiar, menor será a quantidade de características produzidas pelo detector;
2. *nOctaves*: quantidade de oitavas na pirâmide de imagens que o detector usará;
3. *nOctaveLayers*: quantidade de camadas em cada oitava;
4. *extended*: tipo de descritor (vetor com um tamanho de 64 ou 128);
5. *upright*: define a característica de rotação do descritor.

Figura 4.12: Mediana dos pontos de interesse detectados.



O procedimento adotado para a escolha dos parâmetros, consistiu em realizar várias simulações, as quais são identificadas na Tabela 4.2 com o respectivo valor do parâmetro usado. É interessante notar também que os dois últimos parâmetros: *extended* e *upright* são fixos nas simulações, os quais definem o uso de um descritor com 64 posições, e o cálculo da orientação no caso de serem atribuídos o valor *false* para os atributos mencionados.

A Tabela 4.3 apresenta as diversas simulações do algoritmo para a base de dados, contendo as imagens de galáxias do estudo conforme Tabela 3.2 definida no Capítulo 3. A Figura 4.12 apresenta uma visualização das medianas de acordo com a Tabela 4.3. A simulação 3.7 obteve uma maior quantidade pontos de interesse detectados, e terá os valores de seus parâmetros, para construir os histogramas de palavras visuais.

Tabela 4.3: Parâmetros de configuração utilizados no Algoritmo *SURF*.

Simulação	l	o	nO	e	u	E(m)	P(m)
1.1	0	5	3	false	false	79	33
1.2	10	5	3	false	false	58	18
1.3	50	5	3	false	false	19	6
1.4	100	5	3	false	false	9	3
1.5	150	5	3	false	false	6	2
1.6	200	5	3	false	false	5	2
1.7	250	5	3	false	false	4	1
1.8	300	5	3	false	false	3	1
1.9	350	5	3	false	false	3	1
1.10	400	5	3	false	false	3	1
1.11	450	5	3	false	false	3	1
1.12	450	5	3	false	false	2	1
2.1	0	1	3	false	false	77	20
2.2	0	2	3	false	false	79	32
2.3	0	3	3	false	false	79	33
2.4	0	4	3	false	false	79	33
2.5	0	5	3	false	false	79	33
2.6	0	6	3	false	false	79	33
3.1	0	5	2	false	false	76	29
3.2	0	5	3	false	false	79	33
3.3	0	5	4	false	false	81	35
3.4	0	5	5	false	false	81	35
3.5	0	5	6	false	false	81	35
3.6	0	5	7	false	false	82	35
3.7	0	5	8	false	false	83	36
3.8	0	5	20	false	false	83	36

4.7 O uso da técnica *BoF* para mapear as imagens de galáxias aneladas peculiares em histogramas de palavras visuais

Após a extração dos vetores de características dos pontos de interesse, detectados para cada imagem, e das categorias de galáxia do estudado, foi utilizada a técnica *BoF*. Foram realizadas simulações com tamanho do vocabulário de: 32, 64, 128 e 256.

O parâmetro k do algoritmo de agrupamento *k-means* utilizou-se desses valores para a criação dos grupos, e para cada imagem foi elaborado o seu respectivo histograma, de palavra visual baseado na frequência, em que cada vetor de característica aparece em cada um dos grupos, calculados por meio da métrica de distância euclidiana.

Tabela 4.4: Histogramas de palavras visuais.

Categoria	Objeto	C1	C2	C3	C4	C5	C6	...	C32
E	Arp1- 18-chart	7	0	0	4	6	10	...	1
E	Arp1-103-chart	3	0	0	1	0	2	...	1
E	Arp1-255-chart	0	0	0	0	1	0	...	2
P	Arp1-102-chart	2	4	0	0	6	0	...	0
P	Moiseev-38-chart	0	0	0	0	0	0	...	0
P	Arp1- 36-chart	0	0	0	0	1	0	...	0

Tabela 4.5: Histogramas de palavras visuais com as informações das classes.

Categoria	Objeto	C1	C2	C3	C4	C5	C6	...	C32	Classe
E	Arp1- 18-chart	7	0	0	4	6	10	...	1	0,0
E	Arp1-103-chart	3	0	0	1	0	2	...	1	0,0
E	Arp1-255-chart	0	0	0	0	1	0	...	2	0,0
P	Arp1-102-chart	2	4	0	0	6	0	...	0	1,1
P	Moiseev-38-chart	0	0	0	0	0	0	...	0	1,1
P	Arp1- 36-chart	0	0	0	0	1	0	...	0	1,1

A Tabela 4.4 apresenta informações sobre 3 objetos de cada categoria das galáxias de estudo, e a informação do seu histograma de palavra visual criado por meio do *k-means*, considerando o tamanho do vocabulário de dimensão 32. O tamanho 32 do *cluster* foi escolhido por ser a melhor exibição nesse documento, porém o *cluster* de dimensão 256 mostrou uma melhor eficiência conforme apresentado nos testes de similaridade discutidos nas tabelas 4.6, 4.7, 4.8 e 4.9.

Tabela 4.6: Teste de similaridade para a galáxia - 32 - Moiseev-27-chart.

Score	Objeto	Tipo
0.0	Moiseev-34-chart	Polar
0.0	Moiseev-9 -chart	Polar
0.0	Moiseev-27-chart	Polar
1.0	Arp1- 55-chart	Elíptica
1.0	Moiseev-12-chart	Polar
1.0	Arp1-191-chart	Elíptica
1.0	Arp1- 39-chart	Elíptica
1.0	Arp1-185-chart	Elíptica
1.0	Arp1-298-chart	Polar
1.0	Arp1- 77-chart	Elíptica
1.0	Moiseev-13-chart	Polar
1.0	Moiseev-25-chart	Polar
1.0	Arp1- 46-chart	Elíptica
1.0	Moiseev-39-chart	Polar
1.0	Moiseev-36-chart	Polar
1.0	Arp1-110-chart	Elíptica
1.0	Moiseev-23-chart	Polar
1.0	Moiseev-28-chart	Polar
1.0	Moiseev-10-chart	Polar
1.0	Arp1-206-chart	Elíptica

Para a base de galáxias selecionadas para o treinamento foram criados todos os histogramas de palavras visuais totalizando 177 para cada dimensão de vocabulário de 32, 64, 128 e 256, respectivamente.

Ainda nessa fase foram adicionados as informações da classe, sobre a categoria de cada objeto nós diversos histogramas, a exemplo das informações presentes na Tabela 4.5. Para todos os diferentes tamanhos de vocabulários foram criados os histogramas de cada categoria, e adicionado o valor da classe em que o objeto pertence.

Foi realizado o teste de similaridade, por meio da utilização da métrica de distância euclidiana, a qual compara os vetores de atributos para os diferentes *clusters* conforme é apresentado nas tabelas 4.6, 4.7, 4.8 e 4.9 sendo o que obteve resultados mais positivos foi o *cluster* de tamanho 256.

O teste usa uma pesquisa, que tem como entrada uma dada imagem de galáxia, e retorna as imagens similares relacionando-as, por meio de um valor de *score*, que está entre 0 e 10. Quanto mais próximo de 0 estão as imagens, mais parecidas entre si elas serão. Sabendo-se que a galáxia de pesquisa é do tipo polar Moiseev-27-chart, e analisando os valores de retorno próximo a zero, em uma escala de 0 a 10, o *cluster* de tamanho 32, apresentam 8 galáxias do tipo elíptica, e 12 pertencentes à mesma classe do tipo da pesquisa.

Tabela 4.7: Teste de similaridade para a galáxia - 64 - Moiseev-27-chart.

Score	Objeto	Tipo
0.0	Moiseev-27-chart	Polar
1.0	Arp1- 55-chart	Elíptica
1.0	Moiseev-12-chart	Polar
1.0	Arp1-191-chart	Elíptica
1.0	Arp1- 39-chart	Elíptica
1.0	Arp1-185-chart	Elíptica
1.0	Arp1-298-chart	Polar
1.0	Arp1- 77-chart	Elíptica
1.0	Moiseev-13-chart	Polar
1.0	Moiseev-25-chart	Polar
1.0	Arp1- 46-chart	Elíptica
1.0	Arp1- 84-chart	Elíptica
1.0	Moiseev-39-chart	Polar
1.0	Moiseev-36-chart	Polar
1.0	Moiseev-4-chart	Polar
1.0	Arp1-206-chart	Elíptica
1.0	Moiseev-20-chart	Polar
1.0	Moiseev-28-chart	Polar
1.0	Moiseev-10-chart	Polar
1.0	Moiseev-33-chart	Polar

A Tabela 4.7 apresenta a mesma quantidade de objetos recuperados pelo teste de similaridade apresentado na Tabela 4.6 porém recuperando alguns outros objetos próximo ao objeto de pesquisa.

O teste de similaridade apresentado na Tabela 4.8 apresenta melhores resultados comparados com os resultados obtidos com *clusters* de tamanhos 32 e 64, recuperando 5 galáxias do tipo elíptica e 15 pertencentes à mesma classe do tipo da pesquisa, neste caso, de uma galáxia do tipo Polar representada pelo objeto Moiseev-27-chart.

Analisamos os resultados do teste de similaridade com um *cluster* de tamanho 256 (Tabela 4.9) e conseguimos ter resultados superiores até então conhecidos comparados aos resultados do *cluster* de tamanhos 128. Recuperamos com o *cluster* de tamanho 256, 3 galáxias do tipo elíptica e 17 pertencente a mesma classe do tipo da pesquisa que é uma galáxia pertencente da categoria das galáxias do tipo Polares *Moiseev-27-chart*.

Foram efetuadas outras análises dos objetos a partir do teste de similaridade, e optamos por utilizar o *cluster* de tamanho 256, pois foi a que obteve resultados mais satisfatórios. Para representar a camada de entrada da RNA utilizada nesse trabalho serão configurados 256 neurônios e 2 referente a camada de saída para representar a informação da classe.

Tabela 4.8: Teste de similaridade para a galáxia - 128 - Moiseev-27-chart.

Score	Objeto	Tipo
0.0	Arp1- 88-char	Polar
0.0	Moiseev-27-chart	Polar
1.0	Arp1- 55-chart	Elíptica
1.0	Moiseev-12-chart	Polar
1.0	Arp1-217-chart	Polar
1.0	Arp1-191-chart	Elíptica
1.0	Arp1- 77-chart	Elíptica
1.0	Moiseev-13-chart	Polar
1.0	Moiseev-25-chart	Polar
1.0	Arp1- 84-chart	Elíptica
1.0	Moiseev-39-chart	Polar
1.0	Moiseev-36-chart	Polar
1.0	Moiseev-28-chart	Polar
1.0	Moiseev-10-chart	Polar
1.0	Moiseev-33-chart	Polar
1.0	Moiseev-20-chart	Polar
1.0	Moiseev-8-chart	Polar
1.0	Arp1- 56-chart	Elíptica
1.0	Moiseev-15-chart	Polar
1.0	Moiseev-4-chart	Polar

Tabela 4.9: Teste de similaridade para a galáxia - C256 - Moiseev-27-chart.

Score	Objeto	Tipo
0.0	Moiseev-27-chart	Polar
1.0	Arp1-55-chart	Elíptica
1.0	Moiseev-12-chart	Polar
1.0	Arp1-217-chart	Polar
1.0	Arp1-191-chart	Elíptica
1.0	Arp1- 39-chart	Elíptica
1.0	Moiseev-13-chart	Polar
1.0	Moiseev-25-chart	Polar
1.0	Moiseev-39-chart	Polar
1.0	Moiseev-36-chart	Polar
1.0	Moiseev-28-chart	Polar
1.0	Moiseev-10-chart	Polar
1.0	Moiseev-33-chart	Polar
1.0	Moiseev-20-chart	Polar
1.0	Moiseev-15-chart	Polar
1.0	Moiseev-4-chart	Polar
1.0	Moiseev-23-chart	Polar
1.0	Arp1-413-chart	Polar
1.0	Arp1-35-chart	Polar
1.0	Moiseev-22-chart	Polar

Capítulo 5

A Ferramenta *RING−Id* e a sua aplicação para a identificação de galáxias aneladas

O presente capítulo apresenta as principais características, da ferramenta *RING−Id* para a identificação automática de galáxias aneladas peculiares. A aplicação da ferramenta para um conjunto de galáxias aneladas elípticas e polares também é apresentada, em uma primeira etapa com a configuração da rede para o treinamento, e posteriormente para a classificação de uma amostra de galáxias não presente no treinamento. Os valores das métricas obtidos em cada fase também são apresentados.

5.1 Características principais

A ferramenta foi desenvolvida utilizando-se a biblioteca *Encog*[1], a qual é um *framework* de inteligência artificial, que suporta não apenas redes neurais mas também outras áreas da inteligência artificial. Alguns algoritmos de aprendizagem de máquina que a ferramenta faz uso são basicamente: Redes Neurais Artificiais, Máquina de Vetor de Suporte, Programação Genética, Redes Bayesianas. O Framework *Encog* está em desenvolvimento ativo desde 2008 e está disponível para JAVA e .NET (Heaton, 2010). Os elementos básicos de construção da rede neural no *Encog* estão estruturados em camadas, classe lógica neural e funções de ativação.

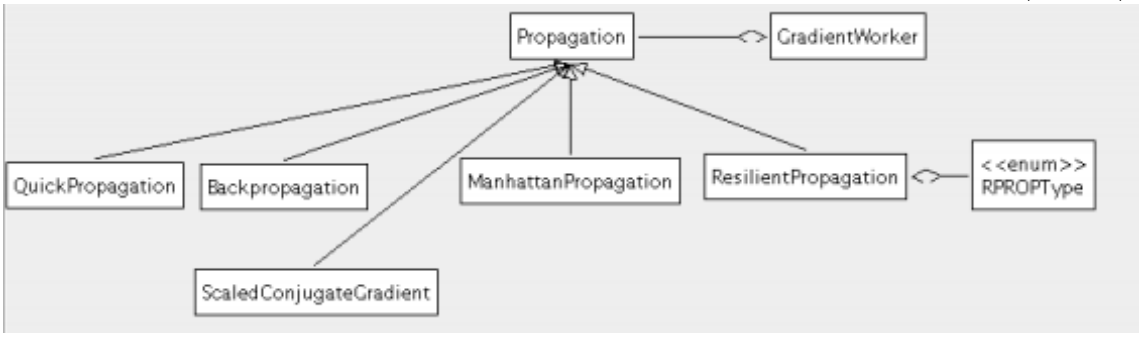
As principais interfaces que implementam a estrutura de uma RNA por meio do *Encog* são a interface para o método de aprendizagem de máquina (classificação, clusterização, regressão), datasets e treinamento (*Backpropagation*, *Manhattan Propagation* e *Resilient Propagation*).

[1]Encog Machine Learning Framework

A Figura 5.1 mostra a interface para a escolha do algoritmo de treinamento da rede. A biblioteca *Encog* suporta os seguintes tipos de algoritmos de retropropagação de erros, os quais são muito comuns para a fase de treinamento em redes neurais artificiais: *Quick Propagation*, *Backpropagation*, *Manhatan Propagation*, *Resilient Propagation* e *Scaled Conjugate Gradient*. Cada tipo de algoritmo mencionado é eficiente para aplicações específicas, que se deseja modelar. A classe base, ou seja, *Propagation* processa os gradientes da rede. A classe *GradientWorker* implementa *multithread* que processa concorrentemente vários tipos de objetos da classe *Propagation*.

Dessa forma, a ferramenta usa uma RNA multicamada com ligações entre camadas *Feedforward* utilizando o algoritmo de treinamento *Backpropagation*, o qual também é conhecido como regra delta generalizada (Banerji et al., 2008). Ainda, segundo os autores, esse tipo de algoritmo é simplesmente um método de gradiente descendente para minimizar o erro quadrático total da saída calculada rede. A ferramenta RING-Id é dividida em dois módulos, o de treinamento e o de classificação da rede, denominados "NDPGTREINO" e "NDPGTESTE", respectivamente.

Figura 5.1: Interface Encog - Algoritmos. Fonte: www.heatonresearch.com/encog/



A configuração da RNA está armazenada no arquivo "NDPG.conf". A Tabela 5.1 mostra com mais detalhes os parâmetros existentes no arquivo de configuração da rede. A seguir, é feita uma breve discussão para cada um desses parâmetros.

O parâmetro *NE* é a entrada da rede neural e definido na aplicação a partir do número de *clusters* representados pelo tamanho do histograma de palavras visuais: 32, 64, 128 e 256, consideramos o *cluster* de tamanho 256, conforme descrito no Capítulo 4, como a quantidade de neurônios para a camada de entrada.

A camada oculta da rede representa pelo parâmetro *NO* faz cálculos intermediários da rede e foi definido com valores de 50, 100, 200 e 400. O parâmetro fixo *NS* definido como 0 ou 1 é a saída da rede neural e é representado na aplicação com dois neurônios. O coeficiente r acelera o aprendizado da rede e foi definido uma faixa padrão de acordo com o existente na literatura, ou seja, 0,2; 0,3 e 0,5.

O parâmetro m não foi utilizado na aplicação, pois não existiu ocorrência de mínimos locais que provocassem uma parada, repentina no momento do treinamento da rede.

O número de épocas definido em intervalos de 200, com valores variando de 200 à 1000, esse parâmetro é utilizado como critério de parada da rede neural. A função de ativação que processa o sinal dos pesos e das entradas, tem a finalidade de gerar um sinal de saída no neurônio. Na ferramenta a função pode ter os seguintes valores:

1. *BiPolar*;
2. *Competitive*;
3. *Linear*;
4. *LOG*;
5. *SIGMOID*;
6. *SOFTMax*;
7. *TANH*.

Tabela 5.1: Parâmetros da Rede.

Nome do parâmetro	Descrição	Faixa
NE	número de neurônios na camada de entrada	32 - 256
NO	número de neurônios na camada oculta	50-400
NS	número de neurônios na camada de saída	2
r	coeficiente de aprendizado	0.2 - 0.5
m	taxa de *momentum*	0.1 - 0.3
EP	número de épocas	200 - 1000
FA	função de ativação da rede	1 - 7

Além dos parâmetros mencionados na Tabela 5.1 o arquivo "NDPG.conf" contém outras informações descritas a seguir:

1. Arquivos de treino e teste: caminho para os arquivos de treinamento e teste;
2. persistência da rede: local do arquivo de persistência. Este arquivo contém todos os pesos finais calculados resultantes do treinamento da rede. Na fase de testes, o algoritmo usa este arquivo de persistência para a classificação.

5.1.1 Requisitos funcionais da ferramenta

A ferramenta para a identificação de galáxias aneladas peculiares foi concebida, para classificar tipos de galáxias aneladas de diferentes categorias. A ferramenta possuí importantes funções, em um processo de classificação de padrões, conforme apresentado a seguir:

1. Leitura dos dados de entrada, validação e de testes a serem utilizados durante os processo de treinamento, validação e de testes da rede;
2. informar a quantidade de neurônios da rede, tais como: quantidade de neurônio nas camadas de entrada, oculta e de saída;
3. informar os atributos: coeficiente de aprendizado, taxa de *momentum*, número de épocas, a escolha da função de ativação;
4. acompanhar a eficiência da rede por meio de arquivos de dados que são usados para visualização gráfica do desempenho da rede;
5. facilitar o processo de avaliação os resultados obtidos por meio da montagem de tabelas e gráficos.

Os dados de entrada, validação e teste foram produzidos pela biblioteca *LIRe*.

5.1.2 Requisitos não funcionais

Considerando o suporte a grandes volumes de dados que a ferramenta irá processar, optou-se em realizar a sua operacionalização com o suporte aos *clusters* de alta performance localizados no Laboratário de Astroinformática (LAi) do INCT-A com sede no IAG-USP. O orientador do trabalho é membro do INCT-A e solicitou uma abertura de conta para o mestrando que realiza os testes e execução do programa. O *cluster* tem a seguinte configuração:

1. O cluster possui 2.304 processadores AMD Opteron 6172 Magny-Cours (12 núcleos) 2.1 GHz/12MB cache;
2. quantidade de memória por *core*: 2 GB;
3. memória por nó de 48 GB;
4. total de memória de 4.6 TB.

O uso desse equipamento com alta performance proporcionou agilidade às atividades. Maiores detalhes podem ser encontrados na página do Laboratório de Informática do IAG-USP[2]. O aplicativo foi compilado na versão do Java 1.7.

[2]https://lai.iag.usp.br

5.1.3 Uma visão geral do módulo de treinamento

O módulo permite que arquivos de treinamento sejam submetidos à rede. Esses arquivos contém os dados para treinamento, assim como as respectivas classes dos tipos de galáxias elípticas e polares. As informações de treino, que serão utilizadas pela rede foram selecionadas de imagens astronômicas do *DSS* descritas no Capítulo 3.

Foram extraídos descritores locais, utilizando a implementação do Algoritmo *SURF* presente na biblioteca *LIRe*. Para as 124 imagens selecionadas para o treinamento da rede, foram elaborados os histogramas, tendo como base os vários tamanhos de vocabulários. Obtivemos o melhor tamanho de vocabulário de 256. A dimensão do vocabulário ideal entre os tamanhos pré-definidos, foi escolhida por meio da análise dos testes de similaridade.

Definido os vários histogramas para cada imagem de galáxia, foram criados arquivos *ASCII* para representar cada imagem. Estes arquivos contém os valores do histograma adicionado aos valores das suas respectivas classes, sendo que a classe polar é representada pelos valores (1,1) e a elíptica pelos valores (0,0). O Capítulo 4 apresenta com mais detalhes como foi realizada a etapa da extração dos descritores, da criação dos diferentes histogramas e dos testes de similaridade para os diferentes tamanhos de vocabulário.

Tabela 5.2: BoF - Formato dos arquivos gerados.

Cluster	Estrutura	Separador	Dimensão	Tamanho
32	numéricos	,	34	1KB
64	numéricos	,	66	1KB
128	numéricos	,	130	1KB
256	numéricos	,	258	1KB

Para uma melhor compreensão dos arquivos gerados, utilizando da técnica *BoF* apresentamos o seu *layout* conforme Tabela 5.2. O *layout* apresenta para cada *cluster* a estrutura, dimensão e tamanho de cada arquivo gerado para cada imagem de galáxia.

Temos como exemplo um *cluster* de tamanho 256, que possui 256 valores numéricos que utiliza o sinal de vírgula "," como separador dos elementos que possui dimensão total de 258. Os primeiros 256 valores da dimensão informada são os dados de entrada para a RNA. Os 2 últimos valores indicam a classe do objeto. O tamanho aproximado de cada arquivo é de 1 kB.

A rede foi submetida a um processo cuidadoso de análise, com o objetivo de escolher o melhor conjunto de parâmetros, tais como a seleção do número de neurônios da camada intermediária, a normalização dos dados de entrada e saída, inicialização

dos pesos da rede, a fixação do coeficiente de aprendizado e da taxa de momento, a seleção da função de transferência. Após os ajustes dos parâmetros, a rede passou por um processo de treinamento. Depois das etapas mencionadas, é esperado como um possível resultado, identificar se uma dada galáxia é do tipo esperado, ou seja, galáxia anelada peculiar polar ou elíptica.

Os pesos da RNA foram inicializados aleatoriamente no intervalo entre 0 e 1. Foram divididos todos os valores de entrada da RNA por um valor que corresponde ao limite máximo das entradas. A normalização é uma estratégia boa quando se usa a função sigmoidal.

Durante o treinamento, a rede procura ajustar os pesos entre as suas conexões, e após ter sido submetida a esta etapa de treinamento, a rede guarda as suas configurações atuais em um arquivo de persistência, para ser utilizado na classificação das novas amostras por meio do módulo de teste.

A RNA atualiza os pesos durante o treinamento, e após a concluir essa fase, os pesos são armazenados em um arquivo *ASCII* para serem utilizados posteriormente na etapa de classificação da rede. A Figura 5.2 apresenta o layout de um arquivo de persistência gerado pela ferramenta com o uso da biblioteca *Encog*. O atributo *weights* armazena todos os pesos finais, ou seja, após o procedimento de treinamento da RNA.

Figura 5.2: Arquivo de persistência gerado após o treinamento da rede.

```
encog,BasicNetwork,java
[BASIC]
[BASIC:NETWORK]
beginTraining=0
layerFeedCounts=2,32,16
weights=0.1431022913,0.4161823701,0.8745284506,0.2878935143,0.4254519134,0.0934726724,0.1012054039,0.5253246798,0.1612590173,
0.2223880825,0.3220969227,0.3787954767,0.7897321649,0.4182944447,0.860243157,0.7452444918,0.6526895025,0.4027072511,0.0622369412,0.0
0.2361204361,0.8545023013,0.9575321993,0.1699231982,0.5979727646,0.7867463315,0.0923377982,0.0376147053,0.9041000493,0.1795287955,0.2
[BASIC:ACTIVATION]
"org.encog.engine.network.activation.ActivationSigmoid"
```

5.1.4 Uma visão geral do módulo de teste

O módulo de teste é utilizado na classificação. Nessa fase, o aplicativo faz a leitura do arquivo de persistência que contém os pesos sinápticos, armazenados pela rede que será a base de conhecimento para a classificação de novas amostras. O tamanho do arquivo e a dimensão dos pesos da rede neste arquivo de persistência dependem da quantidade de neurônios nas camada de entrada, oculta e saída da RNA. A título de exemplo, uma rede com a seguinte configuração: 3 neurônios na camada de entrada, 3 na camada oculta e 2 na camada de saída teria 15 pesos associados às conexões entre camadas.

5.1.5 Módulo de visualização dos dados

A ferramenta produz arquivos no formato texto, os quais são utilizados para obter dados estatísticos das classes, assim como da matriz de confusão. A Figura 5.3 apresenta o formato do arquivo de saída da RNA durante a fase de classificação. O parâmetro *actual* é a resposta predita pela rede e o ideal é o valor correspondente a um tipo de categoria de galáxia, onde (0,0) e (1,1) são as respostas esperadas para uma galáxia da Categoria Elíptica ou Polar, respectivamente.

Figura 5.3: Arquivo de resposta da classificação da RNA.

```
Loading and testing the network
...
0.0 , 0.0 , actual=0.9999395165969558, 0.9999399401222687,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.6414852364482737, 0.6616533315374963,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9999679104457071, 0.9999538737702985,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9833429373952579, 0.99164183263829,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9999327687481571, 0.9999218759149221,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9999456776920385, 0.9999278991465504,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.3095046628250144, 0.3917119760167108,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9990157256283783, 0.9991843471906052,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.0023175244976851033, 0.0017298847712294493,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9993115902497677, 0.9992402623134491,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9977449270499718, 0.9980792619524117,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9962701520857632, 0.9959181647921259,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9997568163605737, 0.9997307773785218,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9998828891222821, 0.9998600447425079,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9867469578703821, 0.9891409104060229,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9899813462910064, 0.9872999512156483,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9995686932809347, 0.9997193838982338,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.999181186842159, 0.9988479093853424,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9566196492581028, 0.9430147452104048,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9911025994333134, 0.9897381105953507,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9999957598220945, 0.999996183043186,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.9998584483243497, 0.9998379213487796,  , ideal=1.0, 1.0
0.0 , 0.0 , actual=0.996619947400421, 0.9953653898657642,  , ideal=1.0, 1.0
0 Hora 0 Minuto 6 Segundos 208 Milisegundo
```

A Figura 5.4 apresenta uma matriz de confusão gerada a partir do arquivo de resposta da classificação representada pela Figura 5.3. A primeira e a segunda linha

representam a classificação correta para os objetos das categorias do tipo elípticas e polares. A quarta e a quinta linha são as classificações incorretas.

Figura 5.4: Matriz de confusão gerada pela resposta da classificação.

```
type_p : 27
type_e : 16
resto : 10
type_e : x3classe_hl= 0 x3classe_cs= 0 x3classe_e= 0 x3classe_p= 6
type_p : x4classe_hl= 0 x4classe_cs= 0 x4classe_e= 4 x4classe_p= 0
```

5.2 A escolha dos melhores parâmetros da ferramenta

A razão para selecionar os melhores parâmetros da aplicação, foi a de encontrar valores adequados para as variáveis, do arquivo de configuração do sistema. Tais valores foram determinados por meio de simulações, que são discutidos posteriormente. Para a camada de entrada da rede, foram utilizados dados dos histogramas de palavras visuais com tamanho 256, conforma já discutido nas seções 4.6 e 4.7.

Para a etapa do treinamento da rede foram inicialmente selecionados um total de 136 galáxias aneladas obtidas dos catálogos FAOA e Moiseev et al. (2011) conforme discutido no Capítulo 3. Com o objetivo de ter-se uma amostra com galáxias, que estavam localizadas em campos sem a presença de galáxias vizinhas e/ou estrelas de *foreground* em um campo de dois minutos de arco, realizamos uma pré-seleção nas imagens, levando em conta os seguintes aspectos:

1. imagens com muitas estrelas de *foreground*;
2. galáxias cujo tamanho eram maiores do que dois minutos de arco (tamanho de nossas imagens).

Após descartar as imagens que não possuíam os dois pré-requisitos acima, o novo conjunto de imagens ficou com 102, 75, 10 e 7 imagens de galáxias aneladas polares, elípticas, Hoag e centralmente suaves, respectivamente.

Dessa forma, a nova amostra contém 194 imagens, as quais foram utilizadas a priori no treinamento da rede. A amostra foi dividida em duas, o conjunto de treinamento e o de testes. A amostra de treinamento tem as seguintes composições: galáxias aneladas polares (71), elípticas (53), *Hoag* (7) e centralmente suaves (5). A amostra de teste é composta por: galáxias aneladas polares (31), elípticas (22), *Hoag* (3) e centralmente suaves (2).

Figura 5.5: Imagens que contém galáxias usadas no treinamento como pertencentes à Categoria Elíptica do Catálogo FAOA. Imagens do *DSS* obtidas por meio do *Aladin*.

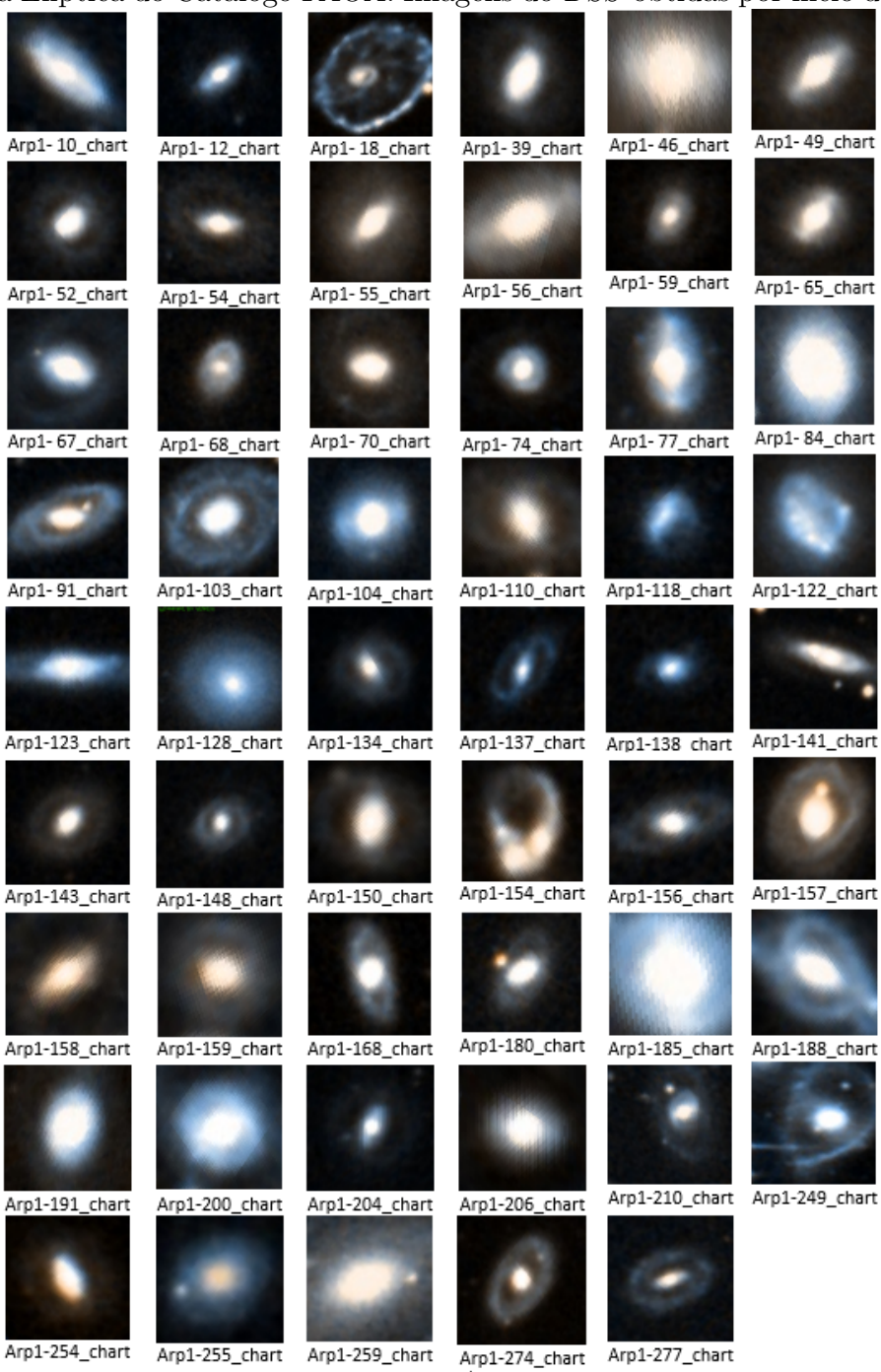
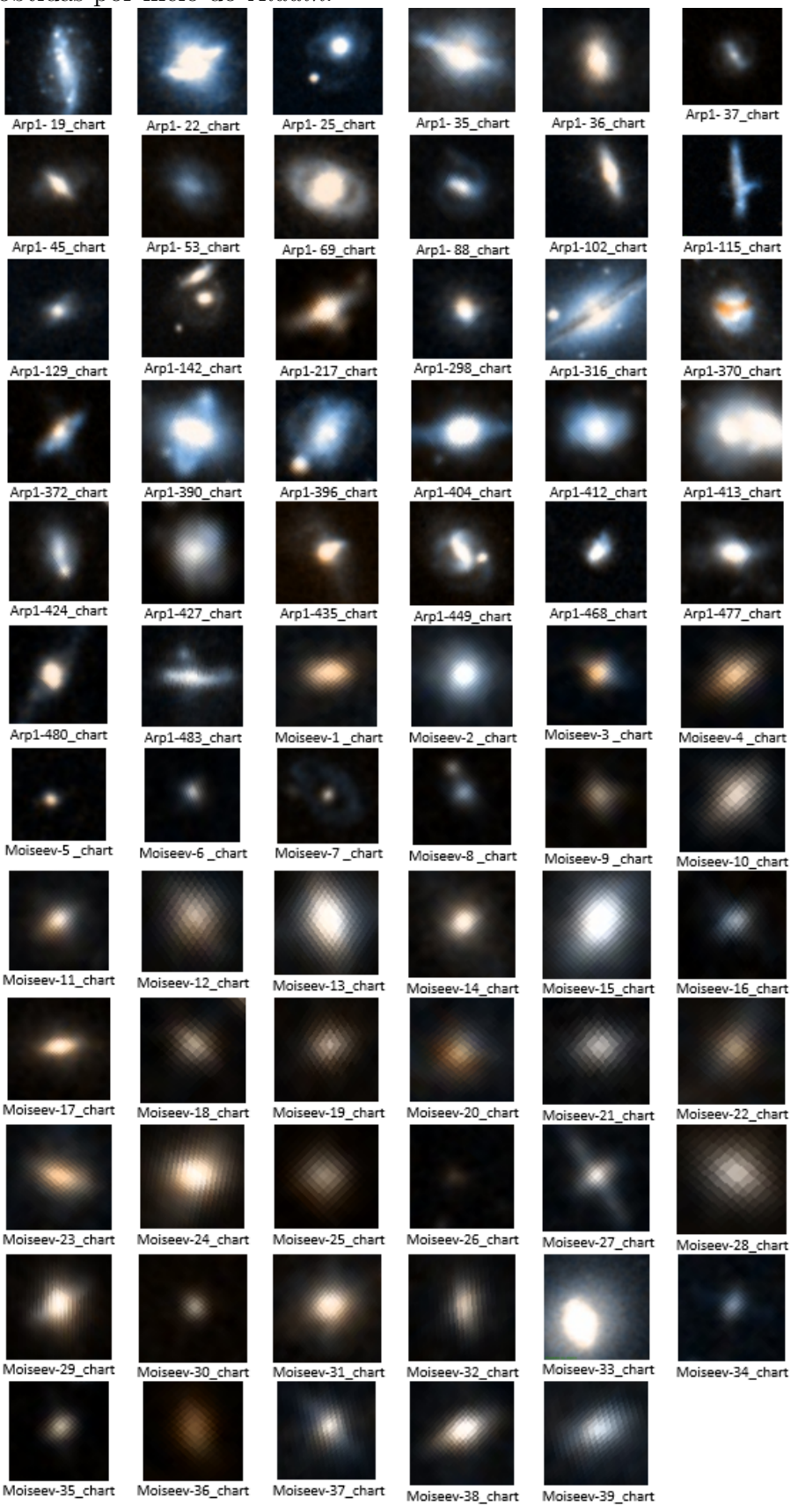


Figura 5.6: Imagens que contém galáxias usadas no treinamento como pertencentes à Categoria aneladas polares dos Catálogos FAOA e Moiseev et al. (2011). Imagens do *DSS* obtidas por meio do *Aladin*.



Usando as quatro categorias mencionadas, nota-se uma classificação ruim para os objetos da Categoria *Hoag* e centralmente suaves. A razão é que após essa seleção, existiam poucas imagens pertencentes a esses tipos de galáxias. Dessa forma, optou-se por descartar a identificação de galáxias de ambas as categorias, utilizando apenas as polares e elípticas, perfazendo um total de 124 e 53, imagens utilizadas no treinamento e na etapa de classificação, respectivamente.

As figuras 5.5 e 5.6 apresentam imagens, que contém galáxias da categoria polares e elípticas utilizadas na amostra de treinamento da RNA. Por sua vez, as galáxias aneladas elípticas apresentam uma forma de um elipsoide. As galáxias do tipo anéis polares são sistemas formados por uma galáxia hospedeira, e estrelas que orbitam em um plano aproximadamente polar.

Um procedimento comum na fase de treinamento da rede, consiste em encontrar o melhor conjunto de parâmetros, tendo em conta as características de nosso problema. A Tabela 5.3 apresenta os valores para as faixas de parâmetros usadas. Alguns valores, a exemplo do coeficiente de aprendizado, estão de acordo com valores usualmente encontrados na literatura (Maren et al., 1990). Cabe ressaltar, que não encontramos entre os melhores resultados de simulação, valores que estivessem na borda inferior ou superior dessa faixa de parâmetros.

Tabela 5.3: Faixa de parâmetros usadas nas simulações para determinar o melhor conjunto de parâmetros.

Nome	Sigla	Faixa
Quantidade de neurônio da camada oculta	*NO*	50, 100,200,300 e 400
Número de épocas	*EP*	200, 400, 600, 800, e 1000
Coeficiente de aprendizado	*r*	0,2; 0,3 e 0,5

Os parâmetros da rede que representam o número de neurônios na camada de entrada *NE*, número de neurônios na camada de saída *NS* e da escolha da função de ativação da rede *FA* são fixos. O parâmetro que indica a taxa de *momentum* ou impulso não foi utilizado, pois a rede neural não encontrou nenhum mínimo local. Conforme já mencionado no Capítulo 4 foram utilizados para os parâmetros *NE* (256 neurônios), *NS* (2 neurônios) e o parâmetro *FA* a função sigmóide.

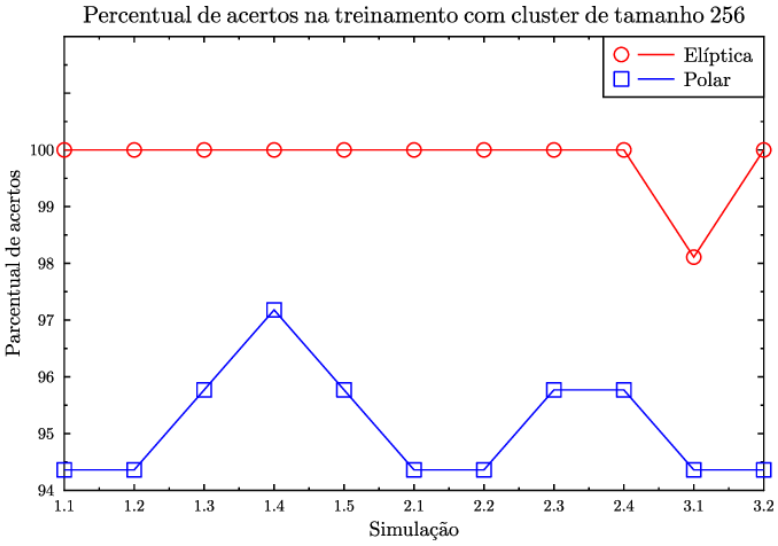
Optou-se por deixar dois parâmetros com valores fixos e alterar o valor de um único parâmetro, esse procedimento foi feito para os três parâmetros. A Tabela 5.4 apresenta os valores utilizados para cada uma das simulações assim como o tempo de processamento, e o respectivo número de acertos para cada uma das duas categorias.

Tabela 5.4: Parâmetros de configuração da rede durante o treinamento para cada simulação.

Simulação	*NO*	*EP*	*r*	segundos	Acertos(E)	Acertos(P)	Geral
1.1	50	200	0,2	27	100%	94,36%	97,18%
1.2	100	200	0,2	48	100%	94,36%	97,18%
1.3	200	200	0,2	88	100%	95,77%	97,88%
1.4	300	200	0,2	156	100%	97,18%	98,59%
1.5	400	200	0,2	244	100%	95,77%	97,88%
2.1	300	400	0,2	309	100%	94,36%	97,18%
2.2	300	600	0,2	462	100%	94,36%	97,18%
2.3	300	800	0,2	706	100%	95,77%	97,88%
2.4	300	1000	0,2	779	100%	95,77%	97,88%
3.1	300	200	0,3	254	98,11%	94,36%	96,23%
3.2	300	200	0,5	250	100%	94,36%	97,18%

Os resultados das simulações podem ser melhor visualizados na Figura 5.7 que apresenta o percentual de acertos para cada simulação. De uma forma geral, nota-se que os índices de acertos são relativamente similares para as simulações.

Figura 5.7: Percentual de acertos no treinamento com *cluster* de tamanho 256.



Na simulação 1.4, o índice de acerto geral foi de 98.59%, o que diferencia essa simulação das demais, é o índice de acerto para galáxias polares, no geral dois pontos percentuais superiores à das demais simulações. Pode também ser visto, que o gasto computacional para essa simulação não é tão grande comparado à das demais simulações do primeiro grupo, i.e., simulações 1.1 a 1.5. Essa simulação considerou um *cluster* de tamanho 256 que foi o melhor valor encontrado nas simulações como abordado no Capítulo 4.

A Figura 5.8 apresenta a evolução do erro com o uso da técnica de validação cruzada durante as épocas de treinamento para a simulação 1.4 (s1.4).

Figura 5.8: Validação cruzada.

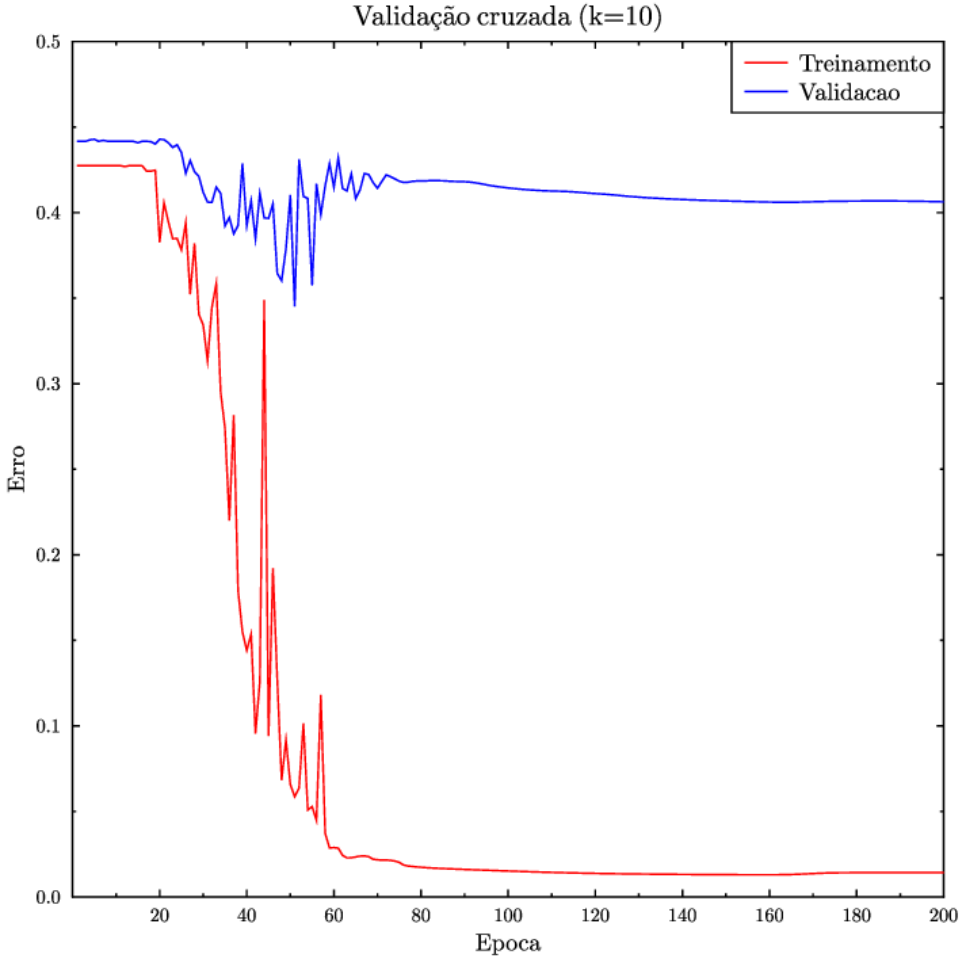


Figura 5.9: Matrizes de confusão gerada pelo processo de treinamento.

(s1.1)

	Polar	Eliptica
Polar	67	4
Eliptica	1	52

(s1.2)

	Polar	Eliptica
Polar	67	4
Eliptica	0	53

(s1.3)

	Polar	Eliptica
Polar	68	3
Eliptica	0	53

(s1.4)

	Polar	Eliptica
Polar	69	2
Eliptica	0	53

(s1.5)

	Polar	Eliptica
Polar	68	3
Eliptica	0	53

(s2.1)

	Polar	Eliptica
Polar	67	4
Eliptica	0	53

(s2.2)

	Polar	Eliptica
Polar	67	0
Eliptica	4	53

(s2.3)

	Polar	Eliptica
Polar	68	3
Eliptica	0	53

(s2.4)

	Polar	Eliptica
Polar	68	3
Eliptica	0	53

(s3.1)

	Polar	Eliptica
Polar	67	4
Eliptica	0	53

(s3.2)

	Polar	Eliptica
Polar	67	4
Eliptica	0	53

Em uma ferramenta de classificação é muito importante analisarmos a matriz de confusão, que é muito útil para a avaliação dos dados. A matriz de confusão possibilita a melhor compreensão nas diversas simulações como apresentado na Figura 5.9.

Figura 5.10: Matriz de Confusão para a simulação 1.4.

(s1.4)

	P	N
P	VP (69)	FN (2)
N	FP (0)	VN (53)

Um exemplo é a simulação 1.4 que apresenta os melhores resultados encontrados. A matriz apresenta a descrição das colunas, que são as classes preditas pelo classificador e as descrições das linhas, na qual é apresentada a classe correta para cada categoria.

A quantidade de objetos presentes em cada categoria são 71 e 53, respectivamente. A matriz dessa simulação para a classe de galáxias tipo anéis polares classifica corretamente 69 objetos e erroneamente 2, classificando-os na categoria anéis elíptica. Para a classe dos anéis do tipo elíptica todos os exemplos são classificados corretamente.

Dada a matriz de confusão para a melhor simulação (s1.4) foram obtidos os valores de precisão e revocação, e logo em seguida foi realizado o cálculo da Medida-F. Detalhes de como são calculados os valores da precisão, revocação e da Medida-F foram discutidos no Capítulo 2.

A classe da amostra de polares da Figura 5.10 foi definida como a principal e a outra secundária. A sua acurácia é dada por:

$$\text{Acurácia} = \frac{VP + VN}{P + N} = \frac{69 + 53}{71 + 53} = \quad 98{,}0\%$$

A métrica da medida de acurácia não é boa, quando o conjunto de dados está muito desbalanceado, podendo facilmente chegar a uma conclusão errônea sobre a sua eficiência.

Para contornar essa questão, penalizamos tanto os falsos positivos quanto os falsos negativos com o uso da métrica da Medida-F. A seguir, são apresentados os valores, de precisão, revocação e Medida-F. Conforme discutido no Capítulo 2, valores próximos de 100% indicam excelentes valores.

$$\text{Precisão} = \frac{VP}{VP + FP} = \frac{69}{69 + 0} = 100{,}0\%$$

$$\text{Revocação} = \frac{VP}{VP + FN} = \frac{69}{69 + 2} = 97{,}0\%$$

$$\text{Medida-F} = 2^* \frac{Precisão * Revocação}{Precisão + Revocação} = 2^* \frac{1 * 0.97}{1.97} = 98{,}6\%$$

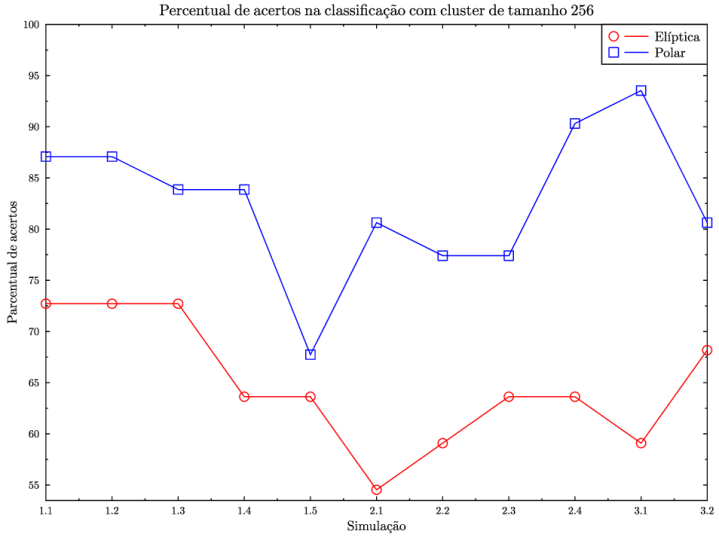
5.3 A classificação da rede

Após o processo de aprendizado da rede e os pesos sinápticos armazenados, foi utilizado o módulo "NDPGTESTE". A amostra para teste contou com 53 objetos para serem utilizados para testar a rede. A amostra contém 22 objetos da categoria elíptica e 31 da polares. O módulo "NDPGTESTE" consultou os pesos sinápticos para cada simulação classificando os objetos na categoria elíptica ou polar. Os resultados do número de acertos para cada uma das simulações são apresentados na Tabela 5.5.

Tabela 5.5: Percentual de acertos esperados por simulações na fase de classificação dos 53 objetos.

Simulação	Segundos	Acertos(E)	Acertos(P)	Geral
1.1	6	72,72 %	87,09%	79,90 %
1.2	6	72,72 %	87,09%	79,90 %
1.3	6	72,72 %	83,87%	78,29 %
1.4	6	63,63 %	83,87%	73,75 %
1.5	6	63,63 %	67,74%	65,68 %
2.1	6	54,54 %	80,64%	67,59 %
2.2	6	59,09 %	77,41%	68,25 %
2.3	6	63,63 %	77,41%	70,53 %
2.4	6	63,63 %	90,32%	76,97 %
3.1	6	59,09 %	93,54%	76,31 %
3.2	6	68,18 %	80,64%	74,41 %

Figura 5.11: Percentual de acertos na classificação com *cluster* de tamanho 256.



A Figura 5.11 apresenta uma visualização dos percentuais para os objetos classificados como corretos de acordo com a Tabela 5.5. Nota-se que para a simulação 1.5 temos o piores resultados para os objetos das duas categorias avaliadas. As simulações 1.1 e 1.2 obtiveram os melhores resultados.

As figuras 5.12, 5.13, 5.14 e 5.15 mostram os resultados da classificação, utilizando os pesos sinápticos armazenados pela simulação 1.1.

Figura 5.12: Objetos classificados corretamente na Categoria Elíptica.

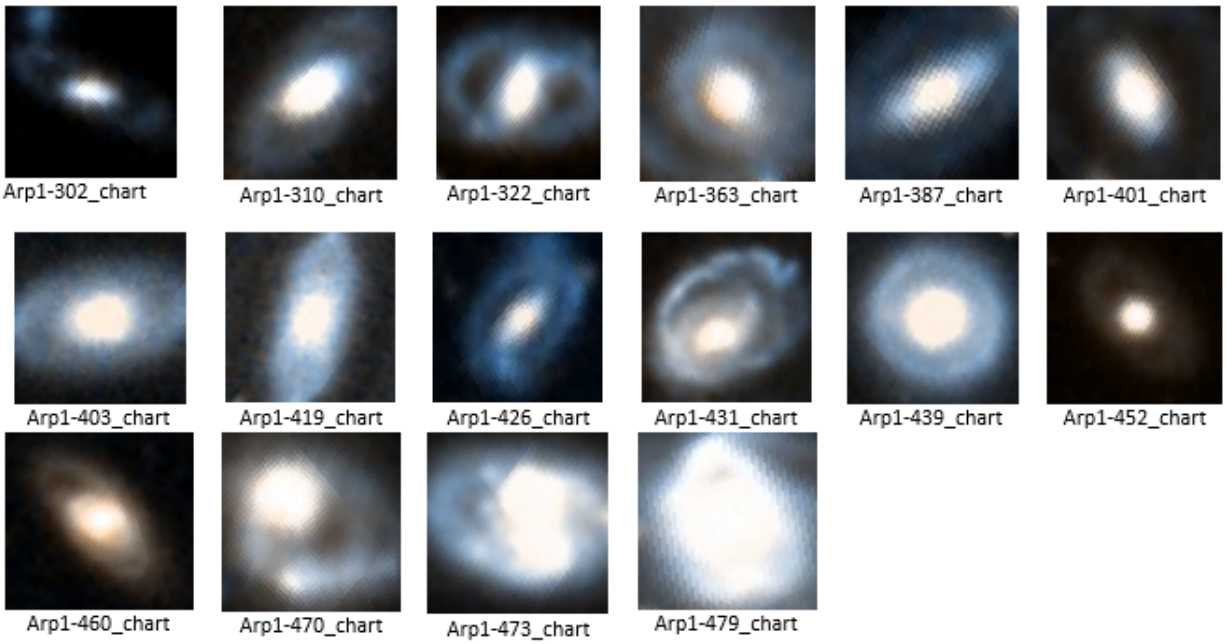
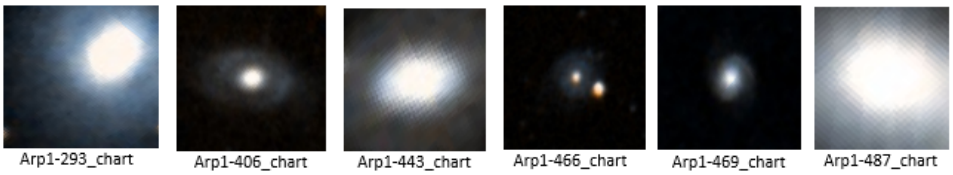


Figura 5.13: Objetos da classe elíptica classificados na Categoria Polar.



As figuras 5.12 e 5.13 apresentam todos os objetos pertencentes a categoria de anéis elípticos classificados corretamente, e dos objetos pertencentes a está classe e classificados erradamente como polares.

Na Figura 5.13 a possibilidade da classificação errada para galáxia de nome Arp1-466_chart, aparentemente é devido a um provável "ruído" próximo da galáxia. A classificação errada para as outras imagens de galáxia presente nesta figura, deve ser pelo motivo da existência de poucos exemplos, dentro do número total de exemplos

de galáxias com aparência próximas as elíticas usadas no treinamento da rede, sendo assim, o sistema não foi capaz de identificá-las como objetos pertencentes a classe da categoria elíptica.

Figura 5.14: Objetos classificados corretamente na Categoria Polar.

Figura 5.15: Objetos da classe polar classificados na Categoria Elíptica.

As figuras 5.14 e 5.15 apresentam todos os objetos pertencentes a categoria de anéis polares classificados corretamente, e dos objetos pertencentes a essa classe e classificados erradamente na categoria anéis elípticos.

Os objetos apresentados na Figura 5.15 o sistema não foi capaz de classifica-lós corretamente. Dois objetos presentes "Moiseev-45_chart" e "Moiseev-56_chart" são aparentemente parecidos com objetos da classe elíptica "Arp1-401_chart" e "Arp1-460_chart".

Figura 5.16: Matrizes de Confusão gerada pelo processo de classificação.

(s1.1)

	Polar	Eliptica
Polar	27	4
Eliptica	6	16

(s1.2)

	Polar	Eliptica
Polar	27	4
Eliptica	6	16

(s1.3)

	Polar	Eliptica
Polar	26	5
Eliptica	6	16

(s1.4)

	Polar	Eliptica
Polar	26	5
Eliptica	8	14

(s1.5)

	Polar	Eliptica
Polar	21	9
Eliptica	9	14

(s2.1)

	Polar	Eliptica
Polar	25	6
Eliptica	10	12

(s2.2)

	Polar	Eliptica
Polar	24	7
Eliptica	9	13

(s2.3)

	Polar	Eliptica
Polar	24	7
Eliptica	8	14

(s2.4)

	Polar	Eliptica
Polar	28	3
Eliptica	8	14

(s3.1)

	Polar	Eliptica
Polar	29	8
Eliptica	3	13

(s3.2)

	Polar	Eliptica
Polar	25	6
Eliptica	7	15

A matriz de confusão gerada durante a fase de classificação é apresentada na Figura 5.16. Como exemplo, pode-se citar as simulações 1.1 e 1.2 que representa os melhores resultados encontrados. A matriz apresenta a descrição das colunas, que representam as classes preditas pelo classificador e as descrições das linhas, na qual é apresentada a classe correta para cada categoria.

A quantidade de objetos presentes em cada categoria são 31 e 22, respectivamente. A matriz dessa simulação, para a classe de galáxias tipo anéis polares, classifica corretamente 27 objetos e erroneamente 4, classificando-os na categoria anéis elíptica. Para a classe dos anéis do tipo elíptica, o método classifica corretamente 16, e erroneamente 6 objetos.

A Figura 5.17 apresenta a simulação 1.1 que contém os melhores resultados encontrados. A matriz apresenta a descrição das colunas, que representam as classes preditas pelo classificador, assim como as descrições das linhas onde é apresentada a classe correta para cada categoria. A quantidade de objetos presentes em cada categoria são 31 e 22, respectivamente.

Figura 5.17: Matriz de confusão para a simulação 1.1.

(s1.1)

	P	N
P	VP (27)	FN (4)
N	FP (6)	VN (16)

A matriz dessa simulação para a classe das galáxias tipo anéis polares classifica corretamente 27 objetos e erroneamente 4, classificando-os erradamente na categoria anéis elíptica. Para a classe dos anéis do tipo elíptica classifica corretamente 16 objetos e erra 6, classificando-os na classe dos anéis polares.

A classe das amostras das polares da Figura 5.17 foi definida como a principal e a outra secundária e em seguida foram calculados os valores da acurácia, precisão, revocação e da Medida-F.

$$\text{Acurácia} = \frac{VP + VN}{P + N} = \frac{27 + 16}{31 + 22} = 81{,}0\%$$

$$\text{Precisão} = \frac{VP}{VP + FP} = \frac{27}{27 + 6} = 63{,}0\%$$

$$\text{Revocação} = \frac{VP}{VP + FN} = \frac{27}{27 + 4} = 87{,}0\%$$

$$\text{Medida-F} = 2^* \frac{Precisão * Revocação}{Precisão + Revocação} = 2^* \frac{0.63 * 0.87}{1.5} = 73{,}0\%$$

A medida de acurácia e a Medida-F apresentam valores percentuais aceitáveis para uma classificação conforme encontrado na literatura (Capítulo 2). O percentual do valor calculado pela métrica de acurácia, não leva em consideração o que é realmente correto ou falso, podendo projetar uma conclusão falsa de uma classificação.

Para contornar o problema são calculados novos valores de precisão e revocação, que penalizam os valores que são falsos positivos e negativos. O valor da medida-F é calculado, a partir da precisão e revocação na tentativa de evitar equívocos no desempenho do algoritmo de classificação.

Capítulo 6

Conclusões e Perspectivas

Apesar dos inúmeros avanços na manipulação e análise de dados de grandes levantamentos astronômicos, ainda existem muitas possibilidades, no tocante não apenas a estudos em grande escala, e na obtenção de parâmetros astrofísicos mas também na identificação de uma vasta gama de tipos de objetos.

Em particular as galáxias, e mais notadamente as galáxias peculiares possibilitam uma grande perspectiva de identificação de novos objetos em dados de grandes levantamentos, alguns ainda não explorados de forma intensa.

No presente trabalho foi elaborado um método, para identificar e classificar objetos como candidatas a galáxias aneladas peculiares das categorias elíptica e polar, presentes nos catálogos de Faúndez-Abans e Oliveira-Abans (1998) e Moiseev et al. (2011). O primeiro catálogo faz parte de uma compilação doe outros catálogos existentes na literatura (ver Capítulo 3) e o segundo do projeto *Galaxy Zoo*.

Analisando as imagens obtidas pelo software *Aladin* para os diferentes catálogos mencionados, observou-se que para algumas imagens, existiam estrelas próximas à galáxias, assim como demais artefatos. Dessa forma, preferiu-se trabalhar com imagens de tamanho de dois minutos de arco, realizando um corte nas mesmas; tendo o ponto base o seu centro, obtido de maneira manual. As imagens obtidas dos diferentes catálogos continham os seguintes tipos de galáxias: Polar (P), *Hoag* (HL), *Elliptical* (E), *Centrally Smooth* (CS).

Optou-se por trabalhar com as categorias dos tipos P e E, pois as outras categorias continham poucas imagens levando o critério acima descrito. A inclusão de outras categorias, poderia levar à uma amostra não representativa do ponto de vista estatístico para essas categorias. Dessa forma, foram elaboradas amostras de treinamento e classificação, com um total de 124 e 53 galáxias, respectivamente para ambas as categorias.

Para a extração de características das imagens, foi usada a biblioteca *LIRe* no software desenvolvido na Dissertação, o $RING{-}Id$. Em um primeiro momento foram

extraídos descritores locais, os chamados pontos de interesses, das diferentes imagens por meio da utilização do Algoritmo *SURF*. Cada ponto encontrado é descrito por um vetor de 64 posições, a partir do qual foi elaborado o histograma de palavra visual para cada imagem.

A *RING−Id* é uma ferramenta que usa a biblioteca de inteligência artificial chamada de Encog. A *RING−Id* foi projetada em dois módulos, o de treinamento da rede e o de classificação, os quais fizeram uso do algoritmo de aprendizado de máquina *Backpropagation*.

A ferramenta ainda permite centralizar em um único arquivo todos os parâmetros de configuração da RNA. Outros resultados produzidos pela ferramenta são armazenados em outros arquivos, como a exemplo do arquivo de pesos da rede e o arquivo de resultados de uma classificação.

O arquivo de resultados gerado pela *RING−Id* proporciona a sua manipulação, para se obter métricas e análise de resultados. Na utilização do Algoritmo SURF e na elaboração dos histogramas foram realizadas diferentes simulações, para a obtenção dos "melhores parâmetros".

A ferramenta *RING−Id* calculou as diferentes entradas dos diversos histogramas, com a mesma dimensão, na fase de treinamento da rede. Dessa forma, foram realizadas diferentes simulações de parâmetros da ferramenta.

Na fase de treinamento da RNA foi alcançado um percentual de acerto para a melhor configuração, de parâmetros de 98,0% e 98,6% para a medida da acurácia e do valor da Medida-F, respectivamente.

Para a fase de classificação da rede o software foi capaz de acessar o valores dos pesos, os quais representaram a melhor configuração na fase de treinamento e classificar amostras desconhecidas até então pela RNA, obtendo os valores de acurácia e da medida-F de 81,0% e 73,0%, respectivamente.

Dentre as possíveis implementações futuras, pretende-se utilizar outros descritores de imagens para tentar melhorar a sua eficiência tanto no treinamento como na classificação. Pretende-se também aplicar método não somente as galáxias aneladas, como também para outros tipos de categorias e em outros grandes levantamentos astronômicos.

Referências Bibliográficas

[Alpaydin 2014] Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

[Amôres et al. 2013] Amôres, E., López-Corredoira, M., González-Fernández, C., Moitinho, A., Minniti, D., e Gurovich, S. (2013). The long bar as seen by the vvv survey-ii. star counts. *Astronomy & Astrophysics*, 559:A11.

[Amôres et al. 2012] Amôres, E., Sodré, L., Minniti, D., Alonso, M., Padilla, N., Gurovich, S., Arsenijevic, V., Tollerud, E., Rodríguez-Ardila, A., Tello, J. D., et al. (2012). Galaxies behind the galactic plane: First results and perspectives from the vvv survey. *The Astronomical Journal*, 144(5):127.

[Arp e Madore 1987] Arp, H. C. e Madore, B. (1987). *A Catalogue of Southern Peculiar Galaxies and Associations: Volume 1, Positions and Descriptions*, volume 1. Cambridge University Press.

[Ball 2001] Ball, N. M. (2001). Morphological classification of galaxies using artificial neural networks. *arXiv preprint astro-ph/0110492*.

[Banerji et al. 2008] Banerji, M., Abdalla, F. B., Lahav, O., e Lin, H. (2008). Photometric redshifts for the dark energy survey and vista and implications for large-scale structure. *Monthly Notices of the Royal Astronomical Society*, 386(3):1219–1233.

[Bay et al. 2008] Bay, H., Ess, A., Tuytelaars, T., e Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

[Bosch et al. 2007] Bosch, A., Muñoz, X., e Martí, R. (2007). Which is the best way to organize/classify images by content? *Image and vision computing*, 25(6):778–791.

[Cerqueira et al. 2016] Cerqueira, S. M. et al. (2016). Identificação de candidatas a galáxias interagentes no infravermelho próximo a baixos redshifts.

[Dieleman et al. 2015] Dieleman, S., Willett, K. W., e Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*, 450(2):1441–1459.

[Dougherty 2009] Dougherty, G. (2009). *Digital image processing for medical applications*. Cambridge University Press.

[Faúndez-Abans et al. 1994] Faúndez-Abans, M., Cuevas, H., e Hertling, G. (1994). Search for faint ring-shaped galaxies in the-77deg to-87deg declination interval. *Astronomy and Astrophysics Supplement Series*, 104.

[Faúndez-Abans e de Oliveira-Abans 1998a] Faúndez-Abans, M. e de Oliveira-Abans, M. (1998a). Looking for fine structures in galaxies. *Astronomy and Astrophysics Supplement Series*, 128(2):289–297.

[Faúndez-Abans e de Oliveira-Abans 1998b] Faúndez-Abans, M. e de Oliveira-Abans, M. (1998b). On the morphology of peculiar ring galaxies. *Astronomy and Astrophysics Supplement Series*, 129(2):357–361.

[Faúndez-Abans et al. 1992] Faúndez-Abans, M., Hertling, G., e Ramírez, A. (1992). The visual appearance of the nuclei of ring-shaped galaxies as an alternative classification criterion. *Astronomy and Astrophysics Supplement Series*, 94:245–250.

[Favan 2015] Favan, J. R. (2015). Utilização de redes neurais artificiais aplicadas na discriminação de padrões de doenças florestais.

[Few e Madore 1986] Few, J. e Madore, B. (1986). Ring galaxies. 2. classification and statistics. *MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY*, 222(4):673.

[Finkelman et al. 2012] Finkelman, I., Funes SJ, J. G., e Brosch, N. (2012). Polar ring galaxies in the galaxy zoo. *Monthly Notices of the Royal Astronomical Society*, 422(3):2386–2398.

[Freitas 2015] Freitas, U. d. P. (2015). Identificação de espécies de peixes utilizando histogramas de palavras visuais em imagens coloridas. Master's thesis.

[Freitas-Lemes et al. 2012] Freitas-Lemes, P., Rodrigues, I., Faúndez-Abans, M., Dors Jr, O., e Fernandes, I. (2012). Imagery and long-slit spectroscopy of the polar ring galaxy am 2020-504. *Monthly Notices of the Royal Astronomical Society*, 427(4):2772–2779.

[Freitas-Lemes 2014] Freitas-Lemes, P. F. L. (2014). *ANLISE ESPECTROFOTOMTRICA DE CANDIDATAS A GALXIAS COM ANEL POLAR*. PhD thesis, Universidade do Vale do Paraíba.

[Gonçalves 2016] Gonçalves, F. M. F. (2016). Uma abordagem interativa guiada por semântica para identificação e recuperação de imagens.

[GONZALEZ e WOODS] GONZALEZ, R. e WOODS, R. Processamento de imagens digitais, 2000. *Editora: Edgard Blücher LTDA*.

[Gonzalez e Woods 2012] Gonzalez, R. C. e Woods, R. E. (2012). Digital image processing.

[Hayati e Shirvany 2007] Hayati, M. e Shirvany, Y. (2007). Artificial neural network approach for short term load forecasting for illam region. *World Academy of Science, Engineering and Technology*, 28:280–284.

[Heaton 2010] Heaton, J. (2010). *Programming neural networks with Encog 2 in Java*. Heaton Research, Inc.

[Ianishi e Izbicki 2017] Ianishi, P. e Izbicki, R. (2017). Classificação morfológica de galáxias em conjuntos de dados desbalanceados. *TEMA (São Carlos)*, 18(1):155–172.

[Jenkinson 2014] Jenkinson, J. (2014). *Enhancement classification of galaxy images*. The University of Texas at San Antonio.

[Kim e Brunner 2016] Kim, E. J. e Brunner, R. J. (2016). Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, pp. stw2672.

[Kovács 2002] Kovács, Z. L. (2002). *Redes neurais artificiais*. Editora Livraria da Fisica.

[Kremer et al. 2017] Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., e Igel, C. (2017). Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*, 32(2):16–22.

[Laganière 2014] Laganière, R. (2014). *OpenCV Computer Vision Application Programming Cookbook Second Edition*. Packt Publishing Ltd.

[Lahav 1996] Lahav, O. (1996). Artificial neural networks as a tool for galaxy classification. *arXiv preprint astro-ph/9612096*.

[Lisin et al. 2005] Lisin, D. A., Mattar, M. A., Blaschko, M. B., Learned-Miller, E. G., e Benfield, M. C. (2005). Combining local and global image features for object class recognition. In *Computer vision and pattern recognition-workshops, 2005. CVPR workshops. IEEE Computer society conference on*, pp. 47–47. IEEE.

[Luger 2005] Luger, G. F. (2005). *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education.

[Madore et al. 2009] Madore, B. F., Nelson, E., e Petrillo, K. (2009). Atlas and catalog of collisional ring galaxies. *The Astrophysical Journal Supplement Series*, 181(2):572.

[Maren et al.] Maren, A., Harston, C., e Pap, R. Handbook of neural computing applications, 1990.

[Moiseev et al. 2011] Moiseev, A. V., Smirnova, K. I., Smirnova, A. A., e Reshetnikov, V. P. (2011). A new catalogue of polar-ring galaxies selected from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 418(1):244–257.

[Muñoz 2009] Muñoz, L. A. B. (2009). Feed-forward artificial neural network basics. In *Encyclopedia of Artificial Intelligence*, pp. 639–646. IGI Global.

[Naim e Lahav 1997] Naim, A. e Lahav, O. (1997). What is a peculiar galaxy? *Monthly Notices of the Royal Astronomical Society*, 286(4):969–978.

[Penatti et al. 2012] Penatti, O. A., Valle, E., e Torres, R. d. S. (2012). Comparative study of global color and texture descriptors for web image retrieval. *Journal of visual communication and image representation*, 23(2):359–380.

[Pham et al. 2005] Pham, D. T., Dimov, S. S., e Nguyen, C. D. (2005). Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1):103–119.

[Sales e Calumby 2010] Sales, B. M. e Calumby, R. T. (2010). Explorando dicionários visuais para recuperação de imagem por conteúdo.

[Shamir e Wallin 2014] Shamir, L. e Wallin, J. (2014). Automatic detection and quantitative assessment of peculiar galaxy pairs in sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 443(4):3528–3537.

[Silva et al. 2010] Silva, I. d., Spatti, D. H., e Flauzino, R. A. (2010). Redes neurais artificiais para engenharia e ciências aplicadas. *São Paulo: Artliber*, pp. 33–111.

[Simon 2001] Simon, H. (2001). Redes neurais–princípios e prática.

[Skrutskie et al. 2006] Skrutskie, M., Cutri, R., Stiening, R., Weinberg, M., Schneider, S., Carpenter, J., Beichman, C., Capps, R., Chester, T., Elias, J., et al. (2006). The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2):1163.

[Storrie-Lombardi et al. 1992] Storrie-Lombardi, M., Lahav, O., Sodre Jr, L., e Storrie-Lombardi, L. (1992). Morphological classification of galaxies by artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 259(1):8P–12P.

[Timmis e Shamir 2017] Timmis, I. e Shamir, L. (2017). A catalog of automatically detected ring galaxy candidates in panstarrs. *The Astrophysical Journal Supplement Series*, 231(1):2.

[Tuccillo et al. 2016] Tuccillo, D., Decencière, E., Velasco-Forero, S., et al. (2016). Deep learning for studies of galaxy morphology. *Proceedings of the International Astronomical Union*, 12(S325):191–196.

[Van Asch 2013] Van Asch, V. (2013). Macro-and micro-averaged evaluation measures [[basic draft]]. *Belgium: CLiPS*.

[Whitmore et al. 1990] Whitmore, B. C., Lucas, R. A., McElroy, D. B., Steiman-Cameron, T. Y., Sackett, P. D., e Olling, R. P. (1990). New observations and a photographic atlas of polar-ring galaxies. *The Astronomical Journal*, 100:1489–1522.