Aplicação de Métodos de Aprendizado de Máquina ao Problema da Separação Estrela/Galáxia

Eduardo Augusto Novo Machado

Projeto de Conclusão de Curso para a graduação em Tecnologia em Sistemas para Internet, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Orientadores:
Eduardo Bezerra, DSc
Ricardo Ogando, DSc

Rio de Janeiro,
Fevereiro 2016

Aplicação de Métodos de Aprendizado de Máquina para o Problema da Separação Estrela/Galáxia

Eduardo Augusto Novo Machado

Orientadores:

Eduardo Bezerra, DSc

Ricardo Ogando, DSc

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M149 Machado, Eduardo Augusto Novo
Aplicação de métodos de aprendizado de máquina ao Problema da Separação Estrela/Galáxia / Eduardo Augusto Novo Machado.— 2016.
v, 39f. : il. (algumas color.) , grafs. , tabs. ; enc.

Projeto Final (Tecnólogo) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2016.
Bibliografia : f. 35-39
Orientador : Eduardo Bezerra
Ricardo Ogando

1. Internet. 2. Internet – Programas de computador. 3. Aprendizado do computador. 4. Algorítmos. 5. Astronomia. I. Bezerra, Eduardo (Orient.). II. Ogando, Ricardo (Orient.). III. Título.

CDD 004.678

Resumo

Para levantamentos astronômicos profundos, recentes ou planejados, é importante separar estrelas e galáxias, uma tarefa que denominamos Problema da Separação Estrela/Galáxia (PSEG). Em magnitudes tênues, a separação entre fontes pontuais e extensas é confusa, o que torna a PSEG uma tarefa difícil. O problema é ainda mais difícil para grandes levantamentos como o Dark Energy Survey (DES) e, num futuro próximo, o Large Synoptic Survey Telescope (LSST) devido aos seus grandes volumes de dados. Sendo assim, a busca por métodos de classificação que são tanto acurados e eficientes é altamente relevante. Neste trabalho, apresentamos uma análise comparativa de vários métodos de aprendizado de máquina com o objetivo de resolver o PSEG em magnitudes tênues. Para treinar os modelos de classificação, o levantamento COSMOS foi utilizado. Usamos métodos de aprendizado de máquina tão distintos como Redes Neurais (Neural Networks (NN)), Support Vector Machines (SVM), k Nearest Neighbors (kNN), Random Forests (RF) e Naive Bayes (NB). O processo exploratório foi modelado como um *workflow* datacêntrico que foi usado para encontrar os melhores valores de parâmetros para cada método de classificação considerado. Os resultados obtidos demonstraram que os métodos NN e RF apresentaram os melhores desempenhos globais em relação às métricas empregadas (acurácia, área ROC, completeza e pureza), recuperando mais de 99% de galáxias e estrelas.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

Capítulo I Introdução

Este capítulo é composto das seguintes seções: Contextualização (seção I.1), Justificativa (seção I.2), Objetivo (seção I.3), Metodologia (seção I.4), Trabalhos Relacionados (seção I.5) e Organização dos Capítulos (seção I.6).

I.1 Contextualização

Na sociedade atual, a tecnologia da informação está presente em quase tudo com que interagimos. A produção de conhecimento está cada vez mais atrelada ao uso do computador e ao avanço do hardware que agiliza a obtenção de resultados. Nesse sentido, nos últimos anos tem ocorrido uma explosão na quantidade de dados gerados e coletados de diversas fontes. Dentre as diversas áreas que fazem uso intensivo de dados para gerar conhecimento, uma das mais fascinantes é a Astronomia. Modernos equipamentos, como a câmera de 570 Megapixel DECam do DES [Flaugher et al., 2015], têm proporcionado uma quantidade enorme de dados, quantidade essa que é impossível de ser tratada sem o auxílio de técnicas computacionais.

Muitos problemas astrofísicos relevantes no que diz respeito à dinâmica das galáxias, efeitos locais na formação e evolução de galáxias e distribuição em grande escala de matéria no Universo são abordados de uma maneira estatística usando levantamentos de galáxias sobre áreas extensas do céu, como o SDSS [Dawson et al., 2016], e DES [Flaugher et al., 2015].

Uma das grandes dificuldades com esses levantamentos é a separação de objetos estelares e não estelares. Particularmente é difícil lidar com objetos tênues (pouco brilhantes) usando até mesmo métodos não computacionais [Mahonen and Frantti, 2000]. Sendo assim, a classificação de objetos astronômicos é de fundamental importância, pois permite um mapeamento correto do Universo, agregando conhecimento às estruturas já conhecidas, sobretudo no que diz respeito a galáxias distantes. Daí a necessidade de ferramentas que permitam a classificação automática de todos os objetos detectados [Cortiglioni et al., 2001].

Com o advento das grandes bases de dados, a necessidade de automatização do processamento de dados tem se tornado urgente. Um perito humano pode classificar objetos com alta precisão baseado na sua aparência em uma placa fotográfica ou imagem digital. Técnicas automatizadas atuais requerem pré-processamento (p. ex. normalização) de dados antes que o dado

bruto seja transformado possibilitando a aplicação de um algoritmo de classificação [Bazell and Peng, 1998].

O problema clássico da análise de repositórios de imagens astronômicas, mais destacadamente em levantamentos extensos do céu, que são agora a maior fonte de dados astronômicos, é a classificação morfológica das fontes detectadas. No nível mais básico, esse problema consiste na separação de fontes entre aquelas não resolvidas espacialmente (estrelas ou possivelmente outros tipos de objetos cujo tamanho angular aparente é menor que a resolução angular efetiva do levantamento, como as galáxias distantes) e as resolvidas espacialmente (galáxias próximas, principalmente). A precisão (objetos de vários tipos corretamente classificados) e a completeza (todos os objetos de um mesmo tipo corretamente classificados) da classificação morfológica da fonte é frequentemente o fator limitante nas aplicações científicas de tais dados, mais do que os limites de detecção [Djorgovski et al., 2006].

A classificação estrela-galáxia diz respeito à tarefa de rotular objetos em uma imagem do céu ou como estrela ou como galáxia, com base em alguns parâmetros extraídos delas [Philip et al., 2002]. Peritos humanos podem usualmente classificar objetos de duas maneiras alternativas: diretamente através de sua aparência numa imagem ou com base nos valores de um conjunto finito desses parâmetros. Porém, a escolha dos parâmetros mais adequados varia grandemente de autor para autor, dificultando a comparação das metodologias [Andreon et al., 2000].

I.2 Justificativa

Até recentemente, tem-se lidado com o problema da separação estrela/galáxia através da análise da morfologia dos objetos. A partir de levantamentos celestes tais como o SDSS passou-se a contar com alternativas a essa abordagem como, por exemplo, a análise das magnitudes ou cores. Porém, a classificação de objetos astronômicos nos limites de detecção é uma tarefa um tanto difícil, mesmo para peritos humanos com habilidades intuitivas e grande experiência.

É então necessário ter máquinas que possam desempenhar a tarefa de classificação com a eficiência de um perito humano (mas com uma velocidade muito mais rápida) e com robustez na classificação sobre variadas condições de observação [Philip et al., 2002]. Técnicas de processamento avançado de imagem e algoritmos de aprendizado de máquina fazem da classificação automatizada de estrelas e galáxias uma alternativa mais rápida à sua contrapartida manual [O'Keefe et al., 2009].

A questão é como analisar esse imenso repositório digital de uma forma sistemática que propicie uma informação útil ao ser humano. Na Ciência da Computação, um campo que tem tido muito desenvolvimento nas últimas décadas para lidar com essa questão é a Mineração de Dados [Odewahn et al., 1992]. A Mineração de Dados entra como auxílio para o pesquisador tirar

suas conclusões, uma vez que automatiza o processo de análise e exibição de informações úteis contidas nos dados. Tais informações podem levar à elaboração de novas teorias que ajudem a explicar como o Universo funciona [Odewahn et al., 1992].

I.3 Objetivo

O presente trabalho tem por objetivo empregar técnicas de Mineração de Dados para revelar padrões ocultos na determinação da natureza de determinados objetos celestes. Neste caso, procurou-se determinar esses padrões através da análise de magnitudes em diversos filtros. Em particular, dado um objeto celeste, desejamos construir um modelo de classificação que permita determinar se se trata de uma estrela ou de uma galáxia. Para abordar esse problema, são empregadas várias métodos de aprendizado de máquina, tais como Redes Neurais (Neural Networks(NN)), Support Vector Machines (SVM), k Nearest Neighbors (kNN), Random Forests (RF) e Naive Bayes (NB), que descrevemos a seguir.

I.4 Metodologia

A solução proposta nesse trabalho é utilizar diversas configurações possíveis na determinação de parâmetros de entrada para funções que implementem as técnicas mencionadas no parágrafo anterior. Como ferramenta de apoio, é utilizada a linguagem “R”, dada a sua grande quantidade de pacotes de aplicação disponíveis, apesar de seu uso pouco frequente em Astronomia. Um *workflow* foi implementado para auxiliar na tarefa. O *workflow* basicamente consiste em atividades de pré-processamento, geração de um modelo através de métodos de aprendizado de máquina, otimização e teste. Os resultados obtidos são avaliados com base na precisão encontrada em um conjunto de dados de controle. A partir daí, a solução otimizada é aplicada a dados observacionais de interesse.

I.5 Trabalhos Relacionados

A pesquisa relacionada ao problema da separação estrela/galáxia usando métodos de aprendizado de máquina ocorre a pelo menos duas décadas. Odewahn et al. [1992] descreveu um experimento usando redes neurais MPL com uma taxa de sucesso de 99% para classificação de galáxias com magnitude aparente $M < 18.5$ e de 95% para $18.5 < M < 19.5$. Outros trabalhos também usaram apenas um método de aprendizado de máquina. Gao et al. [2009] usou random forests e Ball et al. [2006] usou nearest neighbors. Zhang and Zhao [2003] comparou dois métodos de aprendizado de máquina (SVM e Learning Vector Quantization) e observou que SVM teve um desempenho superior a LVQ. Fadely et al. [2012] comparou três métodos (SVM, Hierar-

chical Bayesian e Maximum Likelihood). Em seu experimento, todos os métodos se apresentaram bastante competitivos para classificação de galáxias, retornando 80-90% de completeza por todas as magnitudes. Em termos de pureza, SVM mostrou o melhor desempenho (acima de 90%).

Vários estudos prévios também utilizaram exploração de parâmetros usando um único método de aprendizado de máquina. Podemos destacar o Digitized Sky Survey [Bazell and Peng, 1998] e o Dark Energy Survey [Soumagnac et al., 2015] que exploraram os parâmetros de redes neurais. Adicionalmente, o Sloan Digital Sky Survey (SDSS) [Elting et al., 2008] explorou parâmetros SVM. Entretanto, esses trabalhos não esclarecem a variação dos parâmetros estudados. Vasconcellos et al. [2011] e Zhao and Zhang [2008] exploraram vários tipos de configurações de *decision trees*. Li et al. [2008] usou kNN para explorar diferentes valores de k para encontrar a melhor valor de configuração para a separação entre AGNs e estrela/galáxia. Entretanto, no que concerne ao nosso estudo, nenhum outro trabalho explorou uma ampla variedade de diferentes tipos de métodos de aprendizado de máquina em magnitudes tênues baseado em catálogos astronômicos.

I.6 Organização dos Capítulos

Além desta introdução, este trabalho está organizado nos seguintes capítulos: o Capítulo II apresenta uma visão geral de Astronomia, técnicas de pré processamento de dados e métodos de aprendizado de máquina. O Capítulo III está dividido em seções que abordam os aspectos práticos do trabalho, a metodologia utilizada e os resultados experimentais. Finalmente, o capítulo IV resume os resultados alcançados, tece algumas considerações e aponta para trabalhos futuros.

Capítulo II Fundamentos

Este capítulo é composto das seções de Astronomia (seção II.1), Pré-processamento de Dados (seção II.2) e Métodos de Aprendizado de Máquina para Classificação (seção II.3).

II.1 Astronomia

Neste seção são apresentados alguns conceitos básicos de Astronomia para um melhor entendimento do presente trabalho. Como os objetos astronômicos de estudo são estrelas e galáxias, inicia-se com uma introdução a esses corpos celestes (seções II.1.1 e II.1.2). A seguir, são discutidos alguns aspectos de fotometria e sua utilização na identificação e classificação de objetos em Astronomia (seção II.1.3).

II.1.1 Estrelas

Estrelas podem ser definidas como esferas autogravitantes de gás ionizado cuja fonte de energia é a transformação de elementos através de reações nucleares, da fusão nuclear de hidrogênio em hélio e, posteriormente, em elementos mais pesados até o ferro.

As estrelas tem massas entre 0,08 e 100 vezes a massa do Sol e temperaturas efetivas entre 2500K e 30000K. As estrelas mais massivas que existem são as estrelas azuis com massa de até 100 massas solares e podem chegar a ter uma luminosidade de até um milhão de vezes a do Sol. Por outro lado tem densidades extremamente pequenas. As estrelas mais comuns são estrelas vermelhas (frias) e de baixa luminosidade, chamadas anãs vermelhas. Estas estrelas são muito menores e mais compactas do que o Sol. Uma estrela desse tipo tem massa pequena, em torno de um décimo da massa do Sol, porém, exibem uma densidade em torno de cem vezes a densidade do Sol. No entanto, essas não são as estrelas mais densas que existem. As anãs brancas e as estrelas de neutrons tem densidade muito mais alta. Estrelas de nêutrons em rotação que enviam pulsos de rádio com regularidade são conhecidas como pulsares [de Fátima Oliveira Saraiva, 2004].

Existem ainda estrelas de massa e luminosidade ainda menores, chamadas de anãs marrons, que, por serem muito fracas, são muito difíceis de serem detectadas. Na verdade, anãs marrons são proto-estrelas de massa menor que 0,08 massas solares que nunca queimarão Hidrogênio.

Um tipo especial de estrela são as estrelas variáveis, que são aquelas em que a variação não representa apenas as flutuações normais de grandes conjuntos de partículas em movimentos turbulentos, mas apresentam amplitudes mensuráveis com um certo grau de regularidade.

Algumas estrelas aumentam sua luminosidade rapidamente devido ao início de reações termonucleares descontroladas: as novas e as supernovas. As novas ocorrem em anãs brancas que fazem parte de sistemas binários em que há transferência de massa da companheira para a anã branca. A explosão se dá devido ao acúmulo de matéria transferida que possibilita condições para a queima do Hidrogênio. A curva de luz das novas apresenta um rápido aumento de brilho, da ordem de um dia, seguido de um declínio mais lento. Já as supernovas, muito mais raras, tem aumento de brilho em poucos dias e decréscimo em centenas de dias. São resultado do fim de vida catastrófico de estrelas mais massivas que o Sol.

O diagrama Hertzsprung-Russel, conhecido como diagrama HR, mostra uma relação existente entre a luminosidade de uma estrela e sua temperatura superficial. A maior parte das estrelas está alinhada ao longo de uma estreita faixa na diagonal que vai do extremo superior esquerdo (estrelas quentes e muito luminosas), até o extremo inferior direito (estrelas frias e pouco luminosas). Esta faixa é chamada de sequencia principal (figura II.1). O fator que determina onde uma estrela se localiza na sequencia principal é sua massa: estrelas mais massivas são mais quentes e mais luminosas [de Fátima Oliveira Saraiva, 2004].

II.1.2 Galáxias

Por volta do século XVIII, vários astrônomos já haviam observado entre as estrelas a presença de corpos extensos e difusos, aos quais denominaram “nebulosas”. Hoje sabemos que diferentes tipos de objetos estavam agrupados sob esse termo, muitos pertencendo à nossa própria galáxia. Mas alguns deles, as nebulosas espirais, entre outras, eram galáxias individuais, como a nossa Via Láctea.

Porém, a maioria das nebulosas permanecia com natureza inexplicada. Algumas haviam sido corretamente identificadas como aglomerados estelares e outras como nebulosas gasosas. O problema maior era que a distância até elas não era conhecida, portanto, não era possível saber se pertenciam, ou não, à nossa Galáxia. Posteriormente, usando uma técnica chamada relação período-luminosidade de estrelas variáveis (que variam o seu brilho em intervalos de tempo regulares) de mesmo padrão, independente de sua localização, os astrônomos foram capazes de determinar a distância até as galáxias onde essas estrelas se localizavam.

As galáxias diferem bastante entre si, mas a grande maioria tem formas mais ou menos regulares quando observadas em projeção contra o céu e se enquadram em duas classes gerais: espirais e elípticas. Algumas galáxias não têm forma definida e são chamadas irregulares

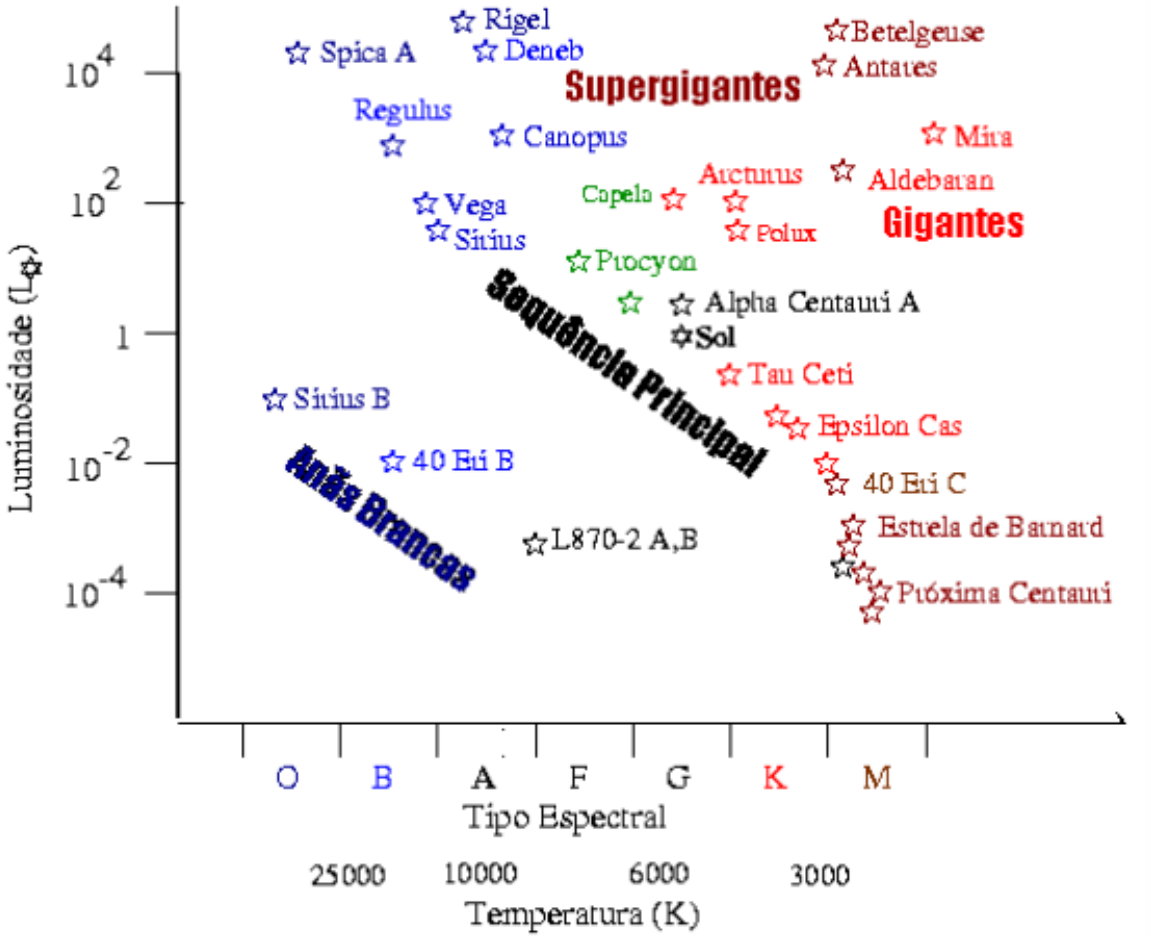

Figura II.1: Luminosidade vs Temperatura

[de Fátima Oliveira Saraiva, 2004].

Olhando-se fotografias do céu, nota-se facilmente que as galáxias tendem a existir em grupos. O grupo de galáxias ao qual a Via Láctea pertence chama-se Grupo Local. Outros aglomerados de galáxias variam de grupos pequenos a grandes aglomerados com centenas de galáxias. Em meados do século passado descobriu-se que aglomerados de galáxias também formam superaglomerados, onde o nosso superaglomerado local é conhecido como Laniakea. Essas estruturas compõe a grande rede cósmica de matéria.

Existem algumas galáxias que emitem uma excepcional quantidade de energia, cuja fonte não são as estrelas. Essas galáxias são classificadas como galáxias ativas e recebem diferentes nomes de acordo com sua aparência e natureza da radiação que emitem (ex: as galáxias Seyfert, as rádio-galáxias e os objetos mais luminosos do Universo - os quasars).

II.1.3 Fotometria

Durante séculos, tem-se procurado quantificar o fluxo e a forma de fontes astronômicas. O fluxo pode variar de acordo com a temperatura e composição química das estrelas, o que revela sua cor. Uma galáxia tem o seu fluxo e cor determinado pelo seu conjunto de estrelas. A cor de uma galáxia também depende de sua formação e pode-se apresentar azulada no caso de galáxias espirais jovens ou avermelhada no caso de galáxias elípticas antigas. Devido ao efeito Doppler e à expansão do Universo [Hubble, 1929], as galáxias distantes tendem a apresentar um desvio para o vermelho, que é maior quanto mais distante a galáxia estiver.

Com o surgimento da fotografia astronômica no fim do século XIX foram aprimoradas as técnicas utilizadas para estudar os corpos celestes. Uma técnica utilizada é a fotometria. Fotometria é a medida da luz proveniente de um objeto. Durante as últimas décadas, muitos tipos de detectores eletrônicos têm sido usados para estudar a radiação eletromagnética. Todo o espectro eletromagnético (faixa de comprimento de onda da radiação eletromagnética que existe no Universo), desde a radiação gama até as ondas de rádio e no visível, é usado para observações astronômicas [de Fátima Oliveira Saraiva, 2004].

Em termos de imagem, a grande diferença entre estrelas e galáxias próximas é que as primeiras geralmente se apresentam como pontuais, ou seja, são não resolvidas. Já as galáxias próximas tendem a se apresentar como uma imagem difusa que se torna menos evidente quanto mais longe e tênue a galáxia esteja localizada.

O brilho aparente de um astro é o fluxo medido na terra (que é a energia por unidade de área e por unidade de tempo que chega ao detector) e, normalmente, é expresso em termos da magnitude aparente que por definição é dada por $m = -2,5logF + const$, onde F é o fluxo e const uma constante que define o ponto zero da escala. A utilização de magnitudes aparentes teve origem com o grego Hiparco que dividiu as estrelas visíveis a olho nu de acordo com seu brilho aparente, atribuindo magnitude 1 à mais brilhante e 6 às mais fracas. No século XIX verificou-se que o sistema, baseado na percepção de brilho do olho humano, é logarítmico, e o fluxo correspondente a uma estrela de primeira magnitude (m=1 mag) era 100 vezes mais brilhante que uma estrela de magnitude 6 (m=6 mag).

Como exemplos de magnitudes aparentes, pode-se citar: o Sol (m=-26 mag), a Lua cheia (m=-12 mag), Sírius (m=-1,4 mag), Marte (m=1,5 mag) e Plutão (m=14 mag). Pode-se notar que a escala é logarítmica invertida, com os astros mais brilhantes tendo valores mais baixos (ou até mesmo negativos).

Em qualquer sistema de magnitudes multicor, define-se os índices de cor como a razão entre os fluxos em duas bandas diferentes. Por exemplo, subtraindo-se a magnitude I (centrada no comprimento de onda 8000Å) da magnitude R (centrada no comprimento de onda 6400Å) temos

o índice de cor R-I e assim por diante. Os índices de cor são importantes para determinar a temperatura das estrelas e tem valores típicos de décimos ou centésimos de magnitudes (pode ser mais, especialmente para galáxias). A figura II.2 mostra as bandas ugriz por comprimento de onda [Fukugita et al., 1996].

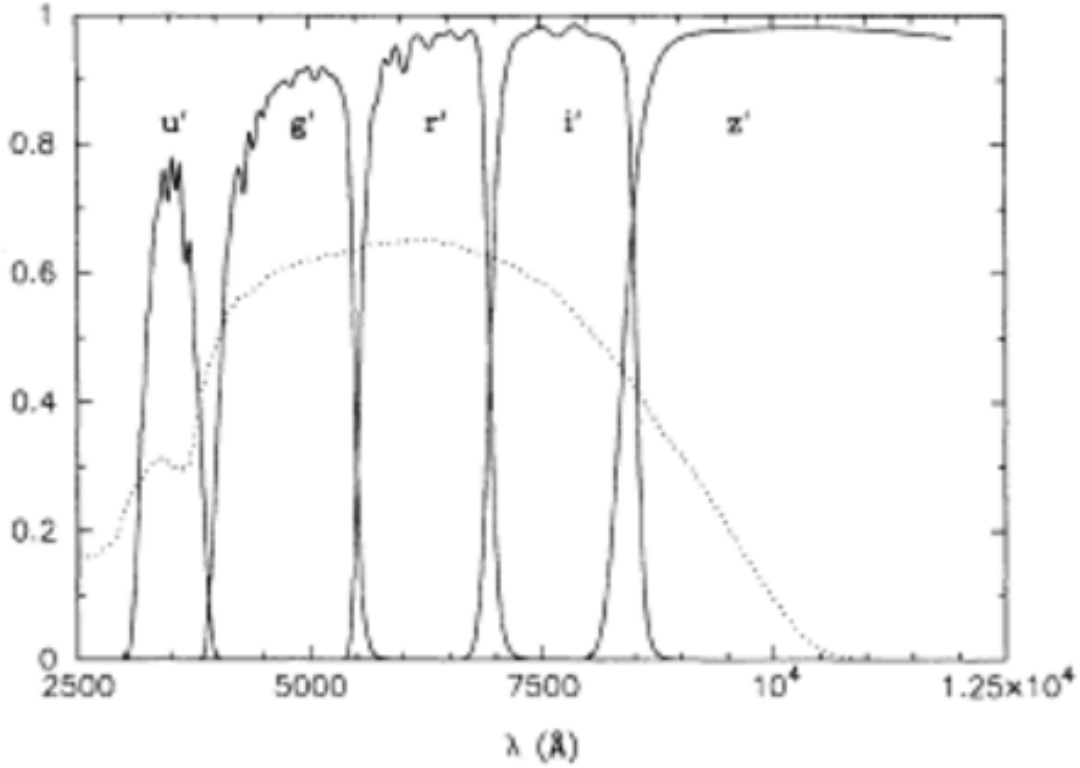

Figura II.2: Bandas ugriz por comprimento de onda

Embora a atmosfera terrestre seja praticamente transparente na faixa visível, ela absorve fortemente no ultravioleta e em várias bandas do infravermelho de modo que não podemos medir ultravioleta no solo e infravermelho somente acima de 2000m de altura. Na atmosfera, existem vários componentes que difundem a luz em todas as direções causando um fenômeno chamado de extinção atmosférica, que afeta todos os comprimentos de onda. Os comprimentos de onda menores são mais absorvidos e espalhados do que os maiores, e, portanto, a luz azul é mais extinguida do que a vermelha. Então, a extinção atmosférica torna as estrelas mais avermelhadas. Sendo assim, as magnitudes observadas devem ser ajustadas aos valores que teríamos se as observações fossem feitas fora da atmosfera. Além da extinção atmosférica, é necessário levar em conta também a extinção interestelar, devido a poeira interestelar concentrada principalmente no plano da galáxia e que também extingue e avermelha a luz das estrelas (é possível corrigir esta extinção usando mapas de poeira, por exemplo, SFD98) [Schlegel et al., 1998].

Em magnitudes aparentes brilhantes relativamente poucas galáxias contaminam um catálogo

de fontes pontuais e relativamente poucas estrelas contaminam um catálogo de fontes resolvidas, fazendo a morfologia ser uma métrica suficiente para a classificação. Entretanto, nos levantamentos celestes da atual geração (como por exemplo o Sloan Digital Sky Survey (SDSS)) e próxima geração (como por exemplo o DES e o LSST), de grande profundidade, existe um vasto número de galáxias não resolvidas em magnitudes aparentes tênues, fazendo com que a classificação de dados seja imprecisa [Fadely et al., 2012].

Quando observamos uma fonte, o fluxo obtido depende da sensibilidade espectral do equipamento, ou seja, do conjunto telescópio + filtro + detector. O filtro é utilizado para escolher a região do espectro eletromagnético que se pretende estudar. Para isso, pode ser empregado um sistema de magnitudes como os sistemas ugriz (SDSS) ou grizY (DES) [Fukugita et al., 1996].

Essas magnitudes são parâmetros básicos utilizadas em levantamentos fotométricos (denominados em inglês *surveys*) como atributos das tabelas de dados e catalogação, junto com seus erros (valores das imprecisões na determinação dessas magnitudes), consistindo em parâmetros fundamentais para os métodos de classificação, assim como parâmetros de forma, elipticidade, etc. [O'Keefe et al., 2009].

Como mencionado acima, o problema começa quando as observações não conseguem mais resolver galáxias e distingui-las de estrelas [Henden and Kaitchuck, 1982; Fadely et al., 2012]. Nas próximas seções, exploramos como podemos usar o pré-processamento de dados e aprendizado de máquina para atacar esta questão.

II.2 Pré-processamento de Dados

Esta seção provê uma visão geral de algumas das principais técnicas de pré-processamento de dados entre as quais a limpeza de dados (seção II.2.1), amostragem (seção II.2.2), remoção de outliers ou valores extremos (seção II.2.3), normalização (seção II.2.4) e criação de folds (seção II.2.5).

Em linhas gerais, Mineração de Dados é o ato de transformar dados observacionais em informação útil. A partir daí, podem ser feitas previsões de caráter hipotético ou teórico. Dada a envergadura dos atuais repositórios de dados, a tarefa de tirar conclusões a partir desses dados necessariamente passa por técnicas de Mineração de Dados. O aumento do poder computacional tem proporcionado o emprego de tais técnicas que já demonstram grande potencial na realização da imensa tarefa da descoberta de conhecimento em base de dados (Knowledge Database Discovery (KDD)), como por exemplo o perfil de compras de uma pessoa. O processo de Mineração de Dados envolve várias etapas. Geralmente, a descoberta de conhecimento a partir de dados envolve algum tipo de pré-processamento [Han et al., 2011].

A tarefa de pré-processamento prepara o dataset original para as atividades subsequentes.

Ela também ajuda a melhorar o desempenho dos métodos de aprendizado de máquina na tarefa de classificação.

II.2.1 Limpeza de Dados

Os dados do mundo real tendem a ser incompletos, apresentando ruídos e erros de medição. Existem algumas técnicas que podem ser utilizadas para resolver o problema dos valores faltantes. Pode-se citar: ignorar as tuplas que tiverem o rótulo de classe faltante, preencher o valor faltante manualmente, usar uma constante global para preencher o valor faltante, usar a média do atributo para preencher o valor faltante, usar a média do atributo para todas as amostras pertencendo a mesma classe de uma dada tupla e usar o valor mais provável para preencher o valor faltante. Entretanto, a ausência de determinado valor em alguns casos não constitui erro.

Outro problema existente em datasets brutos é o ruído. Ruído pode ser definido como um erro aleatório em uma variável medida. Uma técnica que permite remover o ruído é o *binning*, que são métodos que suavizam um valor em uma amostra ordenada consultando sua ”vizinhança”. Estes valores ordenados são distribuídos em um número de *bins* (compartimentos). Pode-se aplicar três métodos para esta técnica, a saber: suavização pela média do *bin*, suavização pela mediana do *bin* e suavização pelo limite do *bin*.

Outra técnica de remoção de ruídos é a regressão, onde dados podem ser suavizados pelo ajustamento a uma função. Pode-se empregar a regressão linear ou a regressão linear múltipla. Finalmente, pode-se mencionar o *clustering*, onde valores similares podem ser organizados em grupos ou *clusters* [Han et al., 2011]

Em Astronomia frequentemente se observa datasets com valores de ruído e erro de medição e é necessário removê-los. Como os datasets geralmente apresentam grande número de registros, pode-se ignorar os registros que apresentam essas inconsistências. A limpeza de dados também é utilizada para gerar um subconjunto do dataset original filtrando através de um critério pré-definido, como por exemplo limitar a faixa de um atributo a determinado intervalo, como uma magnitude limite relacionada ao ruído, que caracteriza a profundidade de um levantamento, que pode ser heterogênea.

II.2.2 Amostragem

Esta atividade consiste em particionar os dados em datasets de treinamento e teste. O dataset de treinamento é usado para gerar o modelo. Após o modelo ser gerado, ele é aplicado ao dataset de teste para gerar previsões para avaliação. A proporção de dados utilizada para particionar os dados de treinamento e teste é variável e usualmente depende da quantidade de dados disponíveis. Um valor típico é de 70/30 podendo chegar a 80/20 para os datasets de treinamento

e teste respectivamente. Para evitar diferenças sistemáticas é normal utilizar um amostragem aleatória sobre o dataset bruto para formar esses datasets.

Os dados do dataset não devem influenciar o modelo gerado. Caso contrário a estimativa de desempenho pode não ser acurada. Deve-se evitar escolher o melhor modelo baseado na repetição de resultados de teste. É conveniente dividir os dados originais em um dataset adicional chamado de validação. Este dataset deve ser usado para iterar e refinar o modelo, reservando o dataset de teste para somente quando um modelo ótimo for gerado. Um valor típico de particionamento é de 60/20/20 para os datasets de treinamento, validação e teste respectivamente.

Uma desvantagem do método que utiliza amostras aleatórias é que os datasets particionados podem não manter a proporção original das classes, ou seja, podem ter maior ou menor proporção de objetos de determinada classe em relação ao dataset original. Em casos extremos, quando uma classe é rara, esta pode ser omitida do dataset de treinamento, levando a uma inconsistência no modelo.

Uma solução para tentar minimizar o problema consiste em empregar a amostra aleatória estratificada. Isto permite que cada dataset particionado pela amostra aleatória estratificada apresente aproximadamente a mesma proporção de objetos de determinada classe do dataset original, por exemplo, se o dataset original possui 90% de galáxias, esta proporção será mantida nos datasets de treinamento e teste. [Lantz, 2013].

O principal problema da amostragem utilizando datasets de treinamento, validação e teste é que grande parte dos dados não pode ser consumida para a geração do modelo, uma vez que pertencem aos datasets de validação e teste, o que pode ocasionar uma piora de desempenho durante a geração do modelo (usando somente o dataset de treinamento). Nesse caso, a técnica de validação cruzada pode ser empregada (seção II.2.5).

Em Astronomia, apesar do DES ser um grande dataset, o COSMOS usado para treinamento e validação não é. O problema fica pior a medida que se explora magnitudes mais fracas. Para assegurar um modelo ótimo baseado em média de desempenho, pode-se usar a validação cruzada no dataset de treinamento conforme mencionado anteriormente.

II.2.3 Remoção de Outliers (Valores Extremos)

Frequentemente encontramos valores de dados que não se enquadram com o comportamento geral ou modelo dos dados. Estes valores são chamados outliers (valores extremos). Outliers podem ser causados por uma variabilidade intrínseca nos dados, ou também por algum tipo de problema nos dados.

A tarefa de remoção de outliers lida com análise de distribuição de dados. Ela remove valores que divergem da distribuição de dados. A remoção de outliers boxplot considera outliers os

valores abaixo de $Q1 - \alpha IQR$ e acima de $Q3 + \alpha IQR$, onde IQR é o *interval quartil range*, $Q1$ e $Q3$ são o primeiro e terceiro quartis, respectivamente. α pode ser definido como 1.5 ou 3.0, dependendo da agressividade ou conservadorismo na percepção de outliers [James et al., 2013].

Em Astronomia, devido a problemas de redução de dados, como pixels quentes ou raios cósmicos que não foram removidos ou mascarados na imagem, as fontes podem ter seu fluxo adulterado, produzindo outliers na distribuição esperada de observáveis.

II.2.4 Normalização

Dois dos métodos mais utilizados na normalização de dados são a normalização min-max e a normalização z-score. A normalização min-max emprega uma transformação linear nos dados originais e preserva o relacionamento entre os valores dos dados originais. Na normalização z-score os valores de um atributo são normalizados baseados na média e desvio padrão do atributo. Este método de normalização é útil quando os valores máximos e mínimos de um atributo são desconhecidos ou quando há outliers que dominam a normalização min-max.

No caso da normalização min-max, os valores se situam entre $-1,0$ e $1,0$ ou entre $-1,0$ e $0,0$. Esta normalização é definida na equação II.1, onde $\hat{x}$ é o valor normalizado, x é o valor original do atributo A, min_A e max_A representam os valores mínimos e máximos do atributo A. $min_{\hat{A}}$ e $max_{\hat{A}}$ são os valores mínimos e máximos desejados para A após a normalização. Desta maneira, o processo de normalização tenta dar pesos iguais a todos os atributos, evitando que atributos com grandes variações sobreponham atributos com menos variação [Ogasawara et al., 2009].

$$\hat{x} = \frac{x - min_A}{max_A - min_A} \cdot (max_{\hat{A}} - min_{\hat{A}}) + min_{\hat{A}} \tag{II.1}$$

Em Astronomia, a normalização de dados é importante pois frequentemente se depara com atributos com intervalos de valores diferentes. Nesse caso, a normalização faz com que estes atributos possam ter seus valores comparáveis uns com os outros, como por exemplo as magnitudes e seus erros.

II.2.5 Criação de Folds

Esta atividade divide os dados em *k* partições chamadas folds. A criação de folds permite o uso de uma técnica chamada validação cruzada com *k*-folds, bastante utilizada para estimar o desempenho do modelo. Estes *k* folds são completamente independentes entre si, ou seja, cada registro do dataset pertence a um único fold, evitando que haja duplicação de registros em cada fold.

O valor mais popular para *k* é dez. Na validação cruzada, para cada um dos dez folds (cada um com um décimo dos dados), um modelo de aprendizado de máquina é construído no restante nove décimos dos dados. Sendo assim, obtem-se dez modelos para avaliar (com dez diferentes combinações de datasets de treinamento e teste). O desempenho global é então avaliado através da média dos dez modelos [Han et al., 2011]

II.3 Métodos de Aprendizado de Máquina para Classificação

Esta seção apresenta alguns métodos de aprendizado de máquina tais como: redes neurais (seção II.3.1), SVM (seção II.3.2), KNN (seção II.3.3), naive Bayes (seção II.3.4) e random forests (seção II.3.5). A seguir são discutidas algumas métricas para avaliação de desempenho (seção II.3.6) e a linguagem utilizada no trabalho (seção II.3.7).

Classificação pode ser descrita como o processo de achar um modelo (ou função) que descreve e distingue classes ou conceitos de dados. O modelo derivado é baseado na análise de um conjunto de dados de treinamento, ou seja, dados para os quais o rótulo de classe é conhecido. O modelo é usado para prever o rótulo da classe que é desconhecido em um outro conjunto de dados com os mesmos atributos (classificação prediz rótulos de classe ou categoria, discretos e não ordenados) [Han et al., 2011].

Classificação é basicamente um processo de dois passos. No primeiro passo constrói-se o modelo de classificação baseado em dados prévios. No segundo passo determina-se se a previsão do modelo é aceitável. Se for, usa-se o modelo para classificar dados novos.

Métodos de aprendizado de máquina para classificação são úteis para automatizar processos de reconhecimento de padrões nos quais a atividade humana é incapaz de lidar com a quantidade de informação analisada.

II.3.1 Redes Neurais

Redes neurais artificiais é uma família de técnicas de inteligência artificial que são capazes de desempenhar tarefas difíceis de classificação de padrões [Odewahn et al., 1992]. Uma rede neural permite considerar um grande número de parâmetros de forma a extrair uma classificação mais acertada e daí ser ideal ao problema que se quer tratar: separação estrela/galáxia num levantamento fotográfico [Odewahn and Nielsen, 1994].

Uma rede neural artificial é um sistema de processamento de informação cuja característica se assemelha àquelas de redes neurais biológicas. O processamento de informação é distribuído entre certo número de unidades simples chamadas neurônios. Através da fase de treinamento são ajustadas as conexões (sinapses) entre esses neurônios e, finalmente, obtem-se um sistema dinâmico que é capaz de reconhecer padrões e classificá-los [Cortiglioni et al., 2001].

Redes neurais quando usadas para classificação é tipicamente uma coleção de unidades de processamento tipo neurônio com conexões pesadas (*weighted*) entre as unidades. Durante a fase de aprendizado a rede aprende ajustando os pesos de forma a ser capaz de prever o rótulo de classe correto das tuplas (conjunto de atributos) de entrada [Han et al., 2011]. Existem muitos tipos diferentes de redes neurais e algoritmos de redes neurais. O algoritmo de rede neural mais popular e amplamente usado em astronomia é o Multi Layer Perceptron (MLP) [Andreon et al., 2002]. MLP requer um conjunto de treinamento cuidadosamente selecionado para uma convergência bem sucedida [Philip et al., 2002]. O algoritmo de MLP com *backpropagation* realiza aprendizado em uma rede neural *feed-forward* multicamada. Esta aprende iterativamente um conjunto de pesos para a predição do rótulo de classe das tuplas. Uma rede neural *feed-forward* multicamada consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída [Han et al., 2011]. A figura II.3 ilustra o processo: cada camada é composta de unidades (1..6). As entradas de rede correspondem aos atributos medidos para cada tupla de treinamento (x1..x3). As entradas são alimentadas simultaneamente nas unidades, constituindo na camada de entrada (1..3). Essas entradas passam através da camada de entrada e são então pesadas (wij) e alimentadas simultaneamente a uma segunda camada chamada de camada oculta (4 e 5). O número de camadas ocultas é arbitrário, embora na prática somente uma é usada. As saídas pesadas da camada oculta (w46 e w56) são entradas para a camada de saída (6), a qual emite a predição de rede para as tuplas dadas.

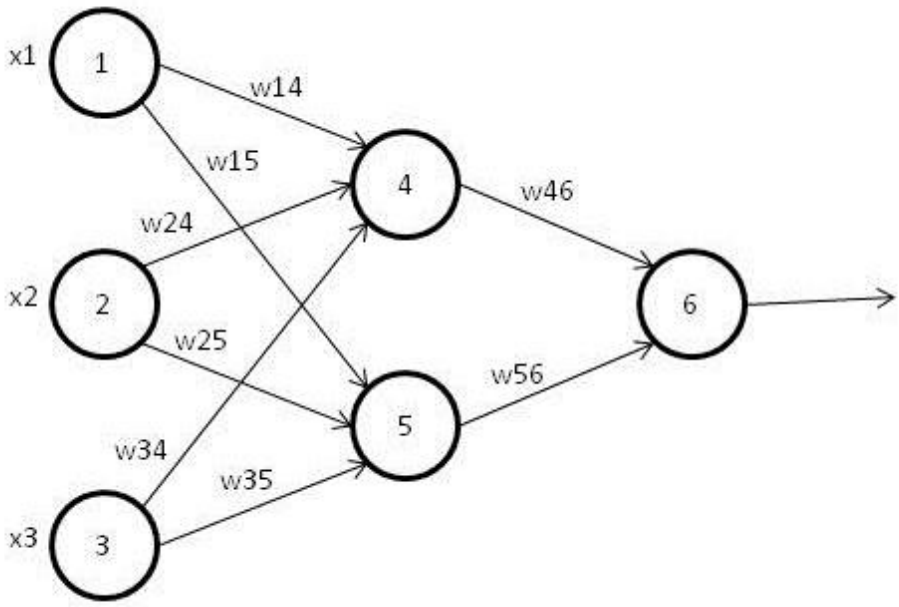

Figura II.3: Exemplo de uma rede neural *multilayer feed forward* [Han et al., 2011]

II.3.2 SVM

Entre outras técnicas de aprendizado de máquina, as máquinas de vetores de suporte (SVM) estão ganhando popularidade em mineração de dados e análise astronômica devido a sua eficácia e relativa simplicidade [Kovács and Szapudi, 2013]. Uma SVM é um tipo de aprendizado de máquina particularmente bem adequado ao problema da classificação [Fadely et al., 2012]. Um conceito central de aprendizado SVM (e de muitos métodos) é o conjunto de treinamento, um conjunto especial de objetos que fornece à maquina exemplos classificados [Kovács and Szapudi, 2013]. A implementação bem sucedida de um algoritmo SVM requer um conjunto de treinamento que seja suficientemente análogo ao conjunto de dados a ser classificado (ou seja, que possuam os mesmos atributos ou colunas de dados) [Fadely et al., 2012].

Em aprendizado de máquina, métodos *kernel* são uma classe de algoritmos para análise de padrões, cujo membro mais conhecido é o SVM. O *kernel* é um algoritmo que realiza um mapeamento não linear para transformar os dados de treinamento originais em uma dimensão mais alta. Dentro desta nova dimensão, o algoritmo procura pelo hiperplano linear de separação ótimo. Com um apropriado mapeamento não linear para uma dimensão suficientemente alta, dados de duas classes podem sempre ser separados por um hiperplano. O SVM acha este hiperplano usando vetores de suporte, isto é, tuplas de treinamento essenciais e margens definidas pelos vetores de suporte [Han et al., 2011]. A figura II.4 ilustra o processo.

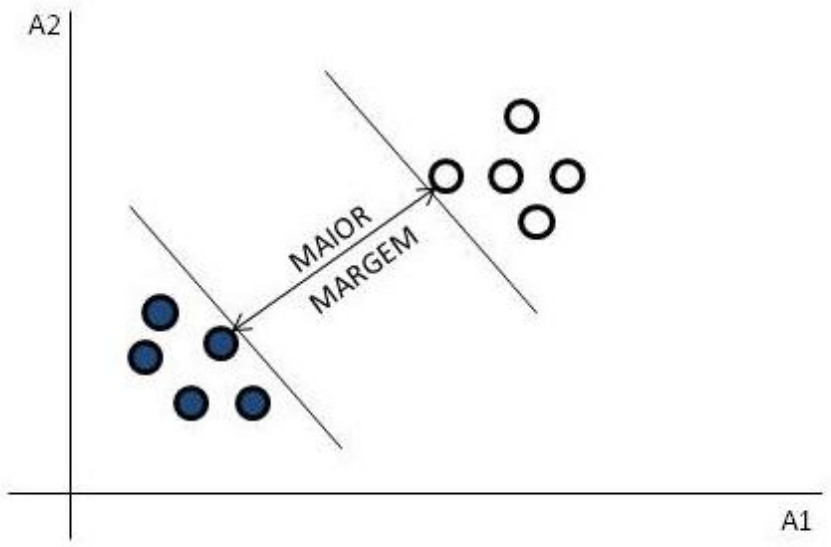

Figura II.4: Vetores de suporte. A SVM acha o hiperplano de separação máximo, isto é, aquele com a máxima distância entre as tuplas de treinamento mais próximas [Han et al., 2011]

II.3.3 kNN

Os métodos de classificação descritos acima são exemplos de *Eager Learners*. *Eager Learners*, quando dado um conjunto de tuplas de treinamento, constroem um modelo de generalização (ou classificação) recebendo novas tuplas (p. ex. de teste) para classificar. Pode-se pensar no modelo aprendido como estando pronto e ávido para classificar tuplas previamente desconhecidas [Han et al., 2011].

Em contraste, métodos *Lazy Learners*, quando dada uma tupla de treinamento, simplesmente a armazenam e esperam até que seja dada uma tupla de teste. Só quando o método vê a tupla de teste, este processa generalização para classificar a tupla baseada em sua similaridade com as tuplas de treinamento armazenadas [Han et al., 2011]. Um dos métodos *Lazy Learners* aqui abordado é o *k-nearest neighbors*.

Em linhas gerais, classificadores kNN são definidos por sua característica de classificar exemplos não rotulados determinando suas classes pela classe dos exemplos rotulados mais similares. O algoritmo *kNN* começa com um conjunto de dados de treinamento consistindo de exemplos que são classificados em diversas categorias, denotadas por uma variável nominal. Deve-se ter um conjunto de dados de teste com exemplos não rotulados com as mesmas características do conjunto de treinamento. Para cada registro no conjunto de teste, *kNN* identifica *k* registros no conjunto de treinamento que são mais "próximos" em similaridade, onde *k* é um inteiro especificado previamente. A instância de teste não rotulado é então determinada pela classe da maioria dos vizinhos mais próximos (*nearest neighbors*) [Bhatia and others, 2010].

Localizar os vizinhos mais próximos requer uma função da distância que mede a similaridade entre duas instâncias. Tradicionalmente, o algoritmo usa a distância euclidiana, que é a distância que se mediria se se pudesse usar uma régua para conectar dois pontos.

Decidir quantos vizinhos são utilizados pelo algoritmo *kNN* determina quão bem o modelo generaliza futuros dados. Escolher um *k* muito grande reduz o impacto da variância causada pelo ruído, mas pode influenciar o aprendizado de tal maneira que corre-se o risco de ignorar pequenos mas importantes padrões [Lantz, 2013].

II.3.4 Naive Bayes

O classificador *Naive Bayes* é um método de classificação baseado no teorema de Bayes. Classificadores Bayesianos são classificadores estatísticos. Eles podem prever as probabilidades de uma determinada tupla pertencer a uma classe particular. Em termos de desempenho, *Naive Bayes* pode ser comparado por exemplo com classificadores usando redes neurais. Além disso, exibem alta acurácia e velocidade quando aplicada a grandes bases de dados.

O algoritmo *Naive Bayes* leva este nome por causa da suposição ingênua (*naive*) sobre os dados, assumindo que todas as características no conjunto de dados são igualmente importantes e independentes. Porém, isto é raramente verdadeiro na maioria das aplicações reais. Entretanto, na maioria dos casos que essas suposições são violadas, *Naive Bayes* ainda se comporta bastante bem [Jiang et al., 2007].

Um problema surge se um evento nunca ocorre para um ou mais níveis de uma classe, ou seja, recebe valor zero. Como as probabilidades em *Naive Bayes* são multiplicadas, este valor zero percentual causa a probabilidade de uma classe acontecer ser zero. Uma solução para este problema envolve utilizar o estimador de Laplace, que essencialmente adiciona um pequeno número às contagens na tabela de frequencias, garantindo que cada característica tenha uma probabilidade não zero de ocorrência em cada classe [Lantz, 2013].

II.3.5 Random Forests

Random Forests é um exemplo de método *ensemble*. Um *ensemble* combina uma série de k modelos de aprendizado (ou classificadores base), $m_1 + \cdots + m_k$, com o objetivo de criar um modelo de classificação composto melhorado m_*. Um dado conjunto de dados D é usado para criar k conjuntos de treinamento $D_1 + \cdots + D_k$, onde D_i (1$\leq$ i$\leq$ k-1) é usado para gerar o classificador m_i. Dada uma nova tupla de dados para classificar, cada classificador base vota em uma determinada classe. O *ensemble* retorna a classificação na classe baseado nos votos dos classificadores base. O *ensemble* tende a ter mais acurácia que os classificadores base [Han et al., 2011].

Random Forests é um método *ensemble* no qual cada um dos classificadores no *ensemble* é um classificador de árvores de decisão, de modo que a coleção de classificadores é uma "floresta". As árvores de decisão individuais são geradas usando uma seleção aleatória de atributos em cada nó para determinar a divisão. Mais formalmente, cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta. Durante a classificação, cada árvore vota e a classe mais popular é escolhida [Verikas et al., 2011].

Random Forests combinam versatilidade e força em um aprendizado de máquina simples. Como o *ensemble* usa só uma pequena e aleatória porção de todas as características do conjunto de dados, este método pode lidar com um conjunto de dados extremamente grande. Ao mesmo tempo, suas taxas de erro para a maioria das tarefas estão niveladas com praticamente qualquer outro modelo [Lantz, 2013].

II.3.6 Métricas para Avaliação de Desempenho de Classificadores

Existe um certo espectro de métricas de avaliação de classificadores. É muito comum em Mineração de Dados medir a acurácia de um classificador [Cortiglioni et al., 2001]. No contexto do PSEG, acurácia (Acc) é definida na equação II.2, onde T_g é o número de galáxias verdadeiras classificadas como galáxias, T_s é o número de estrelas verdadeiras classificadas como estrelas, F_g é o número de estrelas verdadeiras classificadas como galáxias e F_s é o número de galáxias verdadeiras classificadas como estrelas. A acurácia leva em conta a correta classificação de todos os objetos, ao contrário da pureza (ver adiante), que é determinada para cada objeto do dataset.

$$Acc = \frac{T_g + T_s}{T_g + T_s + F_g + F_s} \tag{II.2}$$

Outra medida de desempenho comumente adotada é a área da curva ROC. Uma curva ROC permite visualizar e selecionar o desempenho de um classificador. Para problemas envolvendo duas classes, uma curva ROC permite visualizar o compromisso entre a taxa no qual um modelo M pode reconhecer acuradamente casos positivos (True Positive Rate) versus a taxa no qual ele identifica erroneamente casos negativos como positivos (False Positive Rate) para diferentes porções do dataset de teste. A figura II.5 mostra as curvas ROC de dois modelos de classificação M1 e M2. A linha diagonal representando uma predição aleatória também é mostrada. Assim, quanto mais próxima a curva ROC de um modelo está da linha diagonal, menos preciso é o modelo. Na figura M1 é mais preciso que M2. A área sob a curva ROC é um valor entre 0 e 1. Quanto maior seu valor, maior é a qualidade do classificador correspondente [Fawcett, 2006].

Em Astronomia, a qualidade dos classificadores para o problema da separação estrela/galáxia é também comumente avaliada por duas medidas adicionais: completeza e pureza. Em particular, são funções calculadas de magnitude [Fadely et al., 2012]. Completeza e pureza para galáxias são dadas pelas equações II.3 e II.4, respectivamente. Nestas equações, Cpl_g refere a completeza de galáxias, T_g é o número de galáxias verdadeiras classificadas como galáxias e F_g é o número de estrelas verdadeiras classificadas como galáxias. Pur_g é a pureza das galáxias e F_s é o número de galáxias verdadeiras classificadas como estrelas. Definições similares de completeza e pureza na classificação podem ser feitas para estrelas [Kim et al., 2015].

$$Cpl_g = \frac{T_g}{T_g + F_g} \tag{II.3}$$

$$Pur_g = \frac{T_g}{T_g + F_s} \tag{II.4}$$

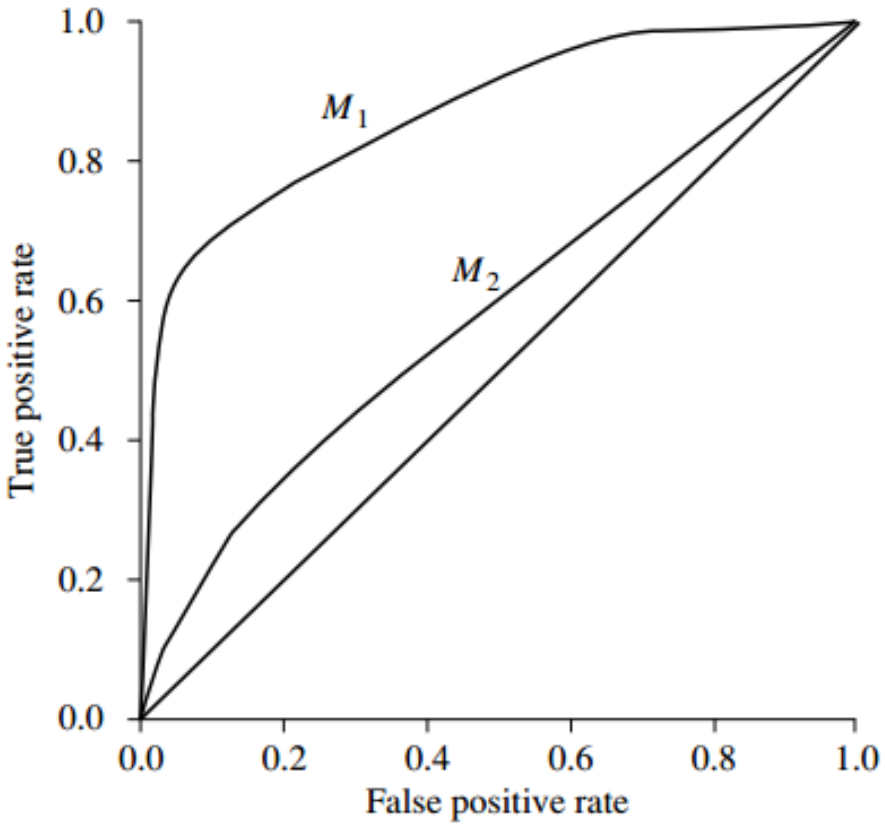

Figura II.5: Exemplo de curva ROC [Han et al., 2011]

É esperado que um método treinado com sucesso alcance altos níveis de completeza e pureza quando aplicados a uma amostra de teste. Entretanto, existem situações onde, para um dado método, pode-se sacrificar completeza pela pureza. Po exemplo, quando executando observações *follow-up*, pode-se querer que a seleção dos alvos seja tão pura (i.e. relevante) quanto possível, evitando desperdiçar tempo de observação do telescópio com fontes mal classificadas.

II.3.7 Linguagem R

A linguagem “R” foi inventada em 1993 e é uma linguagem interpretativa de alto nível que foi originalmente destinada a rodar interativamente quando o usuário gera um comando, obtem um resultado e então gera outro comando. Desde então tem evoluído para uma linguagem que também pode ser embarcada em sistemas e lidar com problemas complexos. Além de transformar e analisar dados, “R” pode produzir gráficos elaborados e relatórios com facilidade. Está sendo usada hoje em dia em todo o seu potencial para análise, extração e transformação de dados, ajuste de modelos, determinação de inferências, obtenção de previsões e representações

gráficas. Seus inúmeros pacotes de aplicação desenvolvidos por uma comunidade de usuários ativa proporciona uma robustez e flexibilidade em uma ampla variedade de áreas onde é utilizada.

Capítulo III Aplicação dos Métodos de Aprendizado de Máquina

Este capítulo é composto da seguintes seções: Metodologia para a Separação Estrela/Galáxia (seção III.1) e Avaliação Experimental (seção III.2).

III.1 Metodologia para a Separação Estrela/Galáxia

Nas seções de metodologia são mostrados alguns aspectos práticos da execução do trabalho como por exemplo a maneira em que foi feito o pré-processamento de dados (seção III.1.1), o uso da validação cruzada (seção III.1.2), a otimização de hiperparâmetros (seção III.1.3) e teste do modelo (seção III.1.4).

A metodologia para a separação estrela/galáxia apresentada neste trabalho é definida como um *workflow* composto por treze atividades exibidas na figura III.1. O *workflow* engloba (A) atividades de pré-processamento (A.1 - seleção e limpeza de dados, A.2 - amostragem, A.3 - remoção de outliers, A.4 - normalização de dados e A.5 - criação de folds para treinamento com validação cruzada), (B) treinamento exploratório com validação cruzada (B.1 - Redes Neurais (NN), B.2 - Random Forests (RF), B.3 - SVM$_{rbf}$, B.4 - SVM$_{poly}$, B.5 - SVM$_{tanh}$, B.6 - kNN e B.7 - Naive Bayes (NB)), (C) Otimização de Hiperparâmetros e (D) Teste.

Antes do *workflow*, a seleção de atributos no dataset COSMOS foi realizada baseada nos filtros do levantamento SDSS [Fukugita et al., 1996]. Isto levou à seleção de dez atributos preditivos e ao rótulo de classe real do objeto. Os primeiros cinco atributos se referem às magnitudes nos filtros *ugriz* [Smith et al., 2002]. No conjunto de cinco filtros *ugriz*, cada letra corresponde aproximadamente a uma diferente banda de magnitude: *u* (ultravioleta), *g* (verde), *r* (vermelho), *i* (infravermelho próximo) e *z* (infravermelho). Os próximos cinco atributos correspondem aos respectivos erros associados a cada magnitude. Nesta primeira abordagem, parâmetros de forma foram deixados de lado. No dataset original, existem três rótulos de classe para cada objeto: galáxia, estrela e quasar.

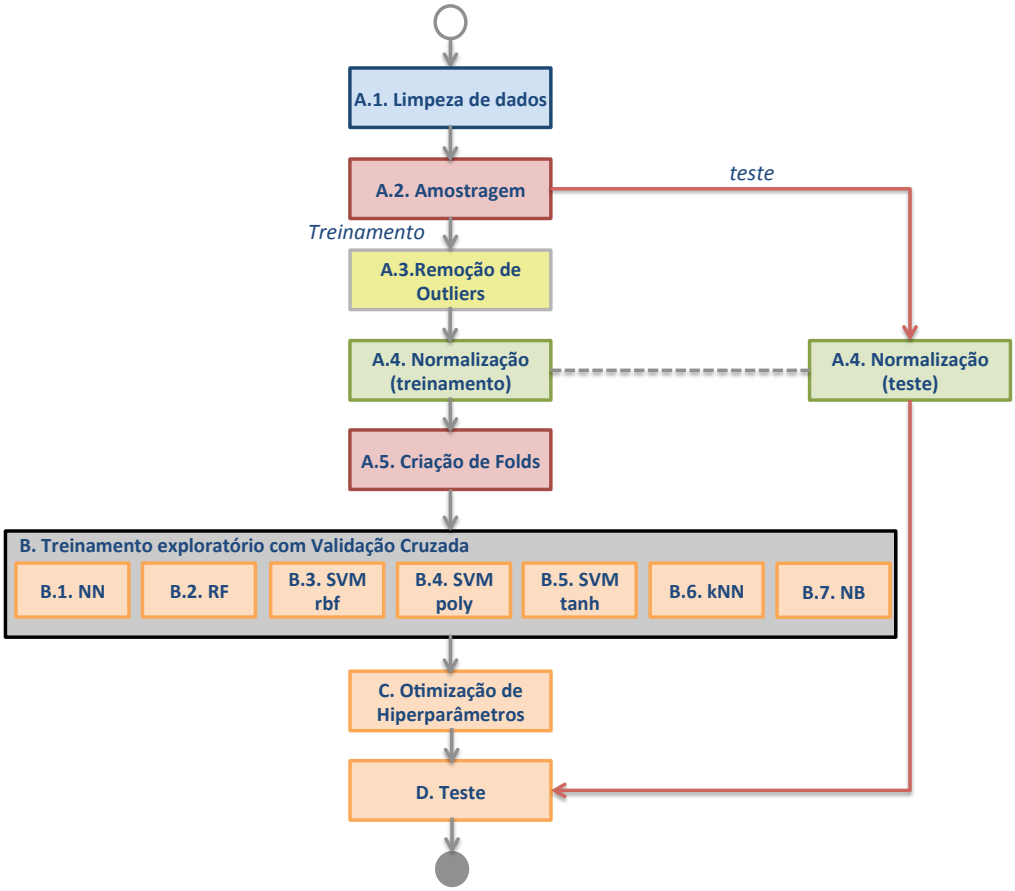

Figura III.1: Workflow de separação estrela/galáxia

III.1.1 Pré-processamento de Dados

Seleção e Limpeza de Dados

Foi feita uma filtragem no dataset COSMOS (ver descrição do COSMOS na seção III.2.1) para selecionar objetos rotulados como estrela ou galáxia e objetos referentes às magnitudes do infravermelho próximo (*i*) variando de 18 a 26 ($18 \leq i \leq 26$). Esta faixa foi escolhida para produzir um dataset comparável com o observado pelo DES.

Adicionalmente, registros contendo valores espúrios nas medidas de telescópio para as magnitudes (anotadas como 99 ou -99 no dataset original) foram descartadas. O dataset passou a contar com 386.957 objetos (5.542 estrelas e 381.415 galáxias)

A tabela III.1 apresenta o efeito desta seleção e limpeza de dados na média ($\overline{x}$) e erro médio ($\overline{\epsilon}$) das colunas de magnitude (*ugriz*). Em particular, valores médios antes e depois da transformação são apresentados. Vale mencionar que a magnitude relacionada ao atributo *i* tem o menor erro de medida entre os outros atributos.

Tabela III.1: Análise dos cinco atributos de magnitudes antes e depois da seleção e limpeza de dados

Magnitude	$\overline{x}_{antes}$	$\overline{\epsilon}_{antes}$	$\overline{x}_{depois}$	$\overline{\epsilon}_{depois}$
u	31.9059	2.46773	25.9895	0.19420
g	28.8307	1.30828	25.6333	0.15136
r	25.6923	0.23712	25.0629	0.09351
i	23.4442	-1.14610	24.6442	0.08907
z	27.6649	1.299191	24.4251	0.17341

Amostragem de Dados

A atividade de amostragem de dados produz uma amostra estratificada de acordo com a descrição apresentada na seção II.2.2. A amostragem divide o dataset em treino/validação (80%) e teste (20%) [Han et al., 2011], produzindo amostras de 309.566 e 77.391 objetos, respectivamente.

Remoção de Outliers (Valores Extremos)

A atividade de remoção de outliers analisa a distribuição de atributos no dataset de treinamento usando critérios *box-plot*. Neste trabalho, valores abaixo de $Q1 - 3 \cdot IQR$ e acima de $Q3 + 3 \cdot IQR$ são considerados outliers e removidos do dataset de treinamento. O resultado dessa remoção produziu uma amostra de 250.745 objetos. A figura III.2 ilustra os resultados antes e depois da remoção de outliers.

Normalização de Dados

Durante a atividade de normalização de dados, foi aplicada a normalização min-max. A normalização foi realizada em todos os atributos de treinamento levando a valores normalizados que variam na faixa $[0, 1]$. Este procedimento de normalização foi usado tanto no conjunto de treino como no de teste.

Criação de k-folds

A atividade de criação de k-folds produz dez folds estratificados para treinamento com validação cruzada [Lantz, 2013]. Estes folds são criados para permitir avaliação exploratória dos métodos de aprendizado de máquina. Todos os dez folds produzidos são usados em todos os treinamentos de aprendizado de máquina para prover condições iguais para todos os métodos avaliados.

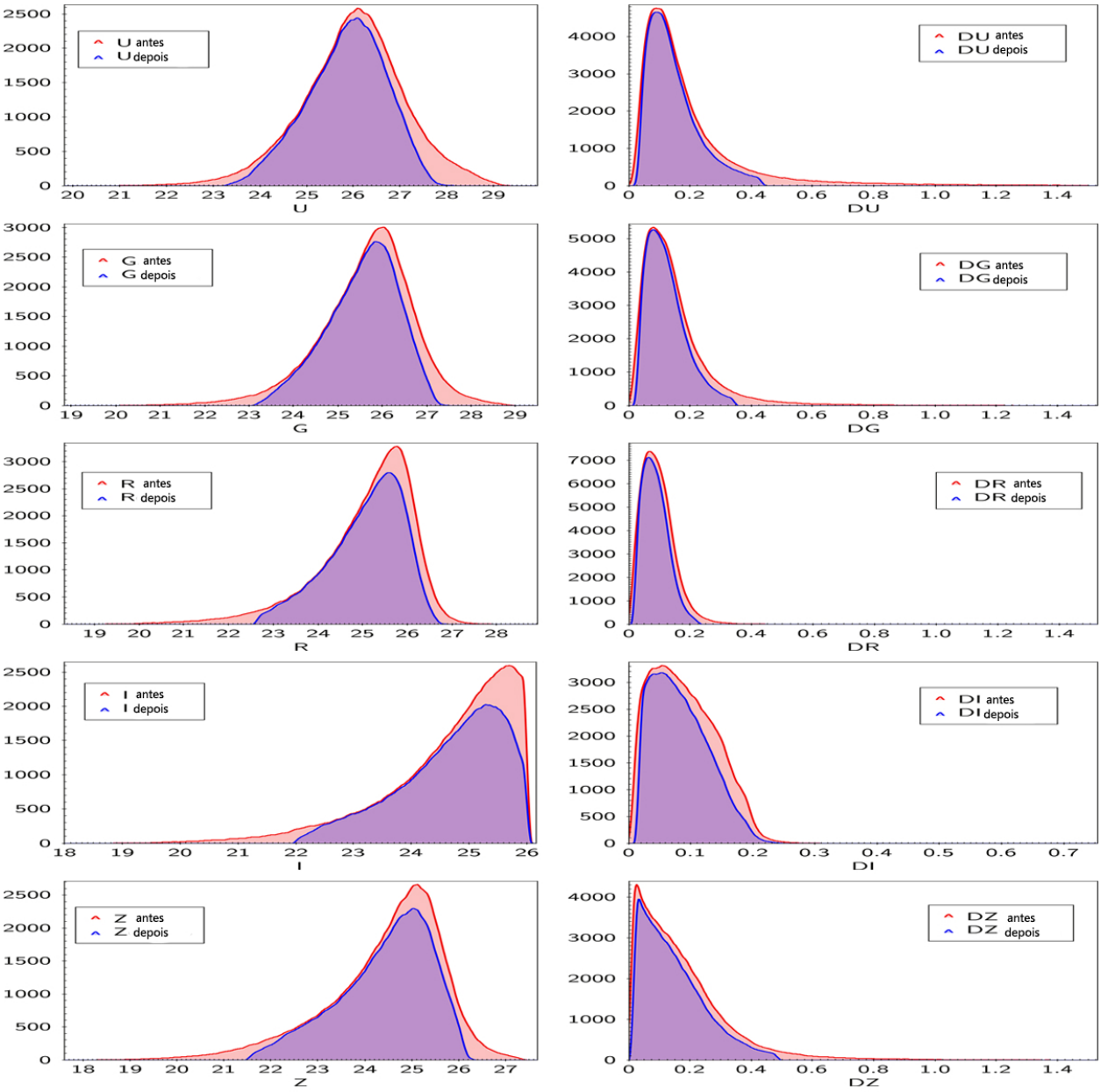

Figura III.2: Distribuição de objetos por magnitudes e seus erros antes e depois da remoção de outliers

III.1.2 Treinamento Exploratório com Validação Cruzada

A partir dos dez folds computados, treinou-se modelos de aprendizado de máquina. Isto resultou na produção de dez diferentes combinações para os conjuntos de treinamento. Cada combinação contendo 90% e 10% para os modelos de treinamento e validação respectivamente, aplicados nos 80% do dataset reservados para treinamento/validação na atividade de amostragem..

Durante a validação cruzada, foram aplicados sete diferentes métodos de classificação, a saber, Redes Neurais Multi-Layer Perceptron (NN), Support Vector Machine (SVM) usando três diferentes *kernels*: rbf, polynomial (poly) e hyperbolic tangent (tanh), k-Nearest Neighbor (kNN),

Random Forest (RF) e Naive Bayes (NB). Todos estes classificadores estão disponíveis como pacotes R (nnet, kernlab, class, randomForest e e1071).

Cada método de aprendizado de máquina explorou um conjunto particular de parâmetros. A tabela III.2 apresenta valores de parâmetros explorados por cada diferente classificador. Como pode ser observado, a quantidade de diferentes configurações de parâmetros foram praticamente os mesmos de modo a permitir uma comparação equilibrada entre eles.

III.1.3 Otimização de Hiperparâmetros

A atividade de otimização de hiperparâmetros objetiva identificar os melhores parâmetros explorados durante o treinamento do modelo com validação cruzada para cada método de aprendizado de máquina [Bergstra et al., 2011]. Para estimar cada modelo de aprendizado de máquina, avaliou-se cada um dez vezes, correspondendo aos dez diferentes folds usados para treinamento com validação cruzada.

A seleção dos parâmetros mais apropriados para cada método foi feita usando a análise de Média-Variância [Wang, 2009]. Procurou-se obter parâmetros que produzem o melhor balanço entre seu desempenho médio e sua respectiva variância. Essencialmente, a análise de Média-Variância leva em conta os resultados da validação cruzada. Para cada configuração de parâmetros, em cada método, é computada a média menos a variância baseada nos dez valores de validação cruzada. Estes valores são determinados como a área da curva ROC. O maior valor entre os resultados de média menos variância aponta a melhor configuração encontrada para cada método. Os melhores parâmetros para cada método de aprendizado de máquina estão indicados na coluna *Melhor* da tabela III.2.

III.1.4 Teste

A atividade de teste consiste em aplicar o modelo de classificação previamente gerado no dataset de teste. Os resultados desta atividade são discutidos no próxima seção.

III.2 Avaliação Experimental

Nas seções da avaliação experimental é descrito o dataset utilizado no trabalho (seção III.2.1) e os resultados obtidos (seção III.2.2).

III.2.1 Descrição do Dataset

O levantamento COSMOS consiste em um catálogo de mais de 500.000 objetos e 90 atributos distribuídos entre de identificação, posição, forma, qualidade, fotométricos, entre outros.

Tabela III.2: Exploração de parâmetros para cada método de aprendizado de máquina

Classificador	Parâmetros	Configuração	Melhor
Neural Network	Neurônios Decay Max. Iterações	3, 4, 5...16 0.01, 0.001 5000	12 0.01 5000
Random Forest	Ntrees	40, 50, 60...300	290
SVM$_{rbf}$	C Sigma	1, 0.2, 0.1 0.1, 0.5, 0.9	1 0.1
SVM$_{poly}$	C Degree	1, 0.2, 0.1 2, 3, 4	1 3
SVM$_{tanh}$	C Scale	1, 0.2, 0.1 0.1, 1, 10	0.1 1
k-Nearest Neighbor	Num Neighbors	1, 2, 3...27	27
Naive Bayes	-	-	-

COSMOS é um levantamento que cobre uma área do céu para o estudo da evolução galática e formação estelar [Scoville et al., 2007], provendo a base para modelos de classificação baseados em catálogo. O importante é que ele tem dados do HST, o qual está fora da atmosfera e permite conhecer bem a natureza das fontes através de sua forma.

Após a limpeza e seleção de dados (ver seção III.1.1), o dataset foi reduzido para 386.957 objetos (5.542 estrelas e 381.415 galáxias). Isto significa que 98.55% do dataset corresponde a galáxias. A figura III.3 representa a densidade de estrelas e galáxias de acordo com a magnitude *i*. Vale notar a alta concentração de estrelas em magnitudes intermediárias. As galáxias exibem um pico entre as magnitudes 25 e 26. Isto significa que um objeto com pouco brilho tem uma maior probabilidade de ser uma galáxia.

A figura III.4 representa a magnitude *i* de acordo com o seu erro de medida. Para magnitudes brilhantes (valores menores que 22), o erro é mínimo. A medida que a magnitude aumenta, o erro tende a se tornar mais evidente para galáxias.

III.2.2 Resultados

A tabela III.3 apresenta o desempenho de diferentes métodos de aprendizado de máquina obtidos durante a exploração de parâmetros [Chirigati et al., 2012] apresentada na tabela III.2. Adicionalmente aos classificadores explorados, foi também incluído o classificador ZeroR, que produz como saída da classificação a classe majoritária presente no dataset, ignorando outras classes [Lantz, 2013]. Em nosso cenário, o ZeroR prediz todos os objetos como galáxias. A importância do classificador ZeroR é indicar um desempenho mínimo para cada método sob es-

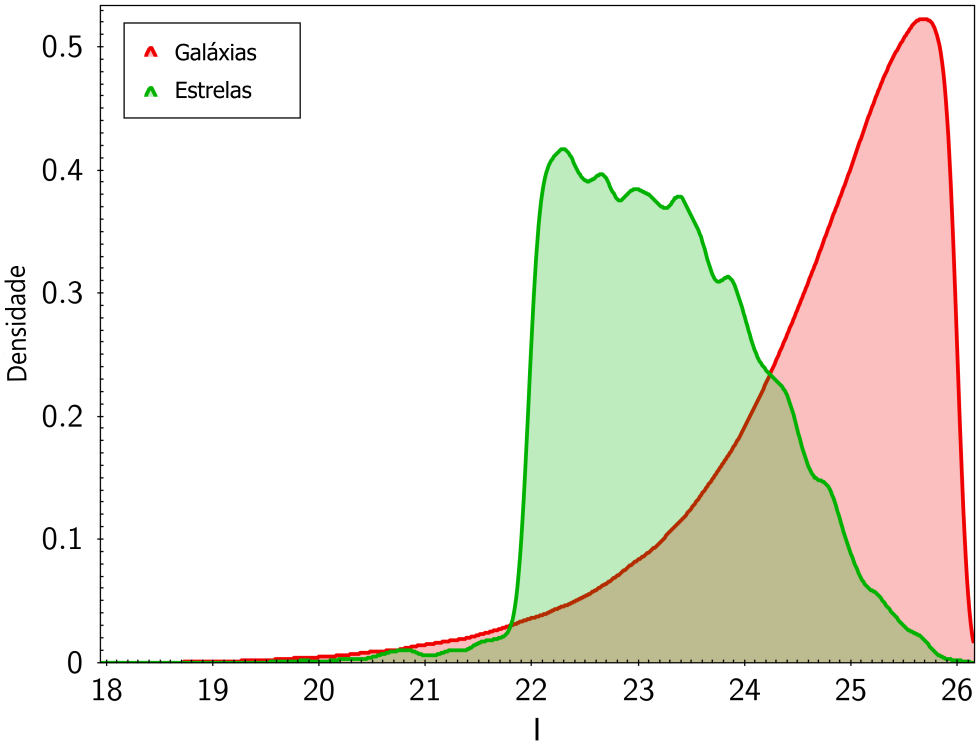

Figura III.3: Magnitude *i* vs densidade de objetos

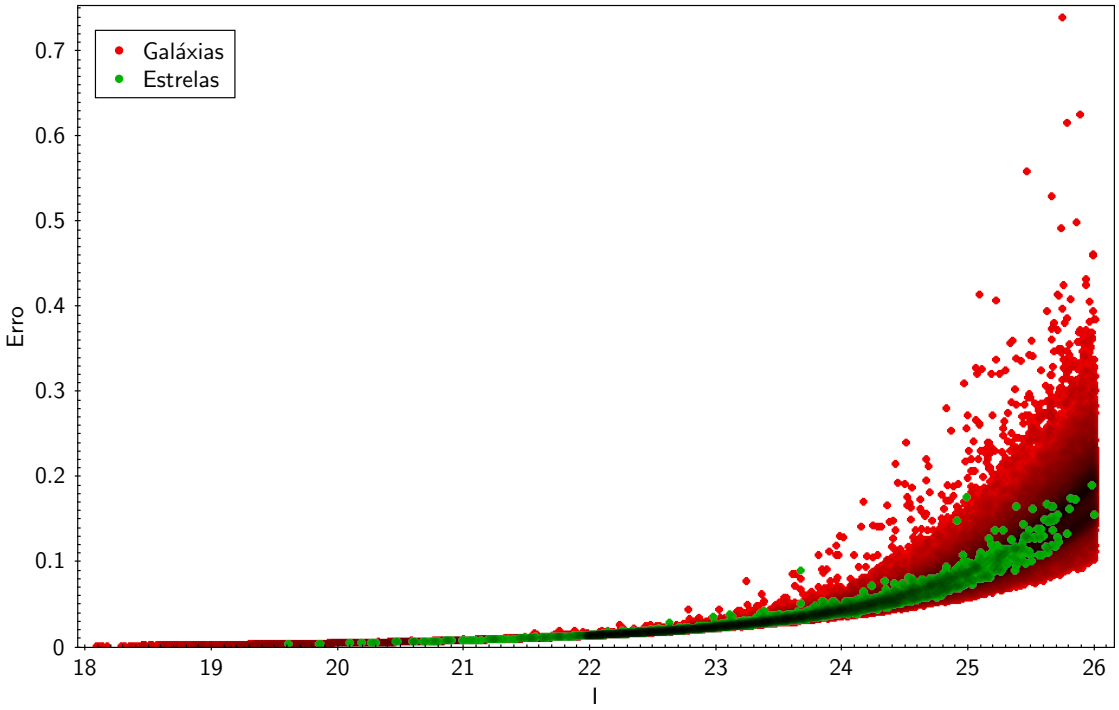

Figura III.4: Magnitude *i* vs seu erro de medida

tudo. Um bom método de classificação deve ao menos mostrar desempenho superior à acurácia do ZeroR, que é de 98,55%.

Foram analisados os resultados de acordo com a acurácia, curvas ROC e completeza e pureza de galáxias. Em relação a medida de acurácia, foi observado que a maioria dos métodos

Tabela III.3: Resultados dos métodos de classificação

Método	Acurácia	AUC	Completeza de galáxias	Pureza de galáxias
NN	**99.19**	**0.984**	99.84	99.34
RF	99.11	0.978	99.87	99.23
SVM$_{rbf}$	99.02	0.913	99.83	99.18
SVM$_{poly}$	98.51	0.961	99.95	98.56
SVM$_{tanh}$	98.55	0.734	**100.00**	98.55
kNN	98.89	0.945	99.87	99.02
NB	83.97	0.869	84.13	**99.54**
ZeroR	98.55	0.734	**100.00**	98.55

tiveram um desempenho superior ao ZeroR. SVM$_{tanh}$ apresentou resultados similares. Entretanto, tanto SVM$_{poly}$ como NB tiveram desempenho pior que o ZeroR. Os métodos NN e RF apresentaram os melhores resultados.

É possível observar da área da curva ROC (coluna AUC - Area Under Curve da tabela III.3) que o método NN teve o melhor desempenho global. NN foi seguido de perto pela RF. O melhor desempenho para método SVM foi produzido usando o *kernel* polynomial. SVM$_{rbf}$, kNN e NB foram melhores que ZeroR, mas com um desempenho relativo inferior quando comparado a NN. Finalmente, SVM$_{tanh}$ teve desempenho insatisfatório.

Analisando completeza e pureza, o cenário ideal seria atingir 100% em ambas as medidas. Entretanto, em situações reais isto raramente é alcançado. Portanto, um bom modelo de classificação deve manter um balanço entre essa duas medidas. Este foi exatamente o caso de NN e RF. De maneira oposta, SVM$_{tanh}$, NB e ZeroR apresentaram um desempenho desbalanceado.

Para melhor entender os resultados, analisamos o desempenho de cada método de aprendizado de máquina para diferentes bandas de magnitude *i* com o modelo já treinado. A figura III.5 apresenta as faixas de magnitude *i* pela pureza para cada classificador. Foi observado que NB apresentou uma anomalia na classificação em magnitudes brilhantes (magnitude *i* entre 18 e 23). Nesta faixa, NB classificou erroneamente todas as galáxias como estrelas. Assim para propósitos de apresentação, a curva NB nesta figura apresenta os resultados somente para magnitudes superiores a 23. Para valores de magnitude superiores a 20, os resultados de todos os classificadores começam a mostrar uma queda de desempenho, que pode ser devido a um acréscimo da proporção de estrelas presentes no dataset. Como se pode notar, a faixa de magnitude entre 22 e 24 exibiu os mais baixos valores de pureza para os classificadores. Todos os métodos, exceto RN (NN), mostraram uma abrupta queda de desempenho dentro dessa faixa.

Isto pode ser devido ao fato de que a maioria das estrelas do dataset estão concentradas nessa faixa (figura III.3). Considerando todos os objetos presentes nessa faixa, 10% deles são estrelas. Isto claramente modifica o balanceamento global do dataset, que é em torno de 1,5%. Para magnitudes maiores que 24, o desempenho dos classificadores torna a aumentar, uma vez que a proporção de estrelas comparada a de galáxias diminui. Além disso, o SVM tanh apresentou um comportamento similar ao ZeroR.

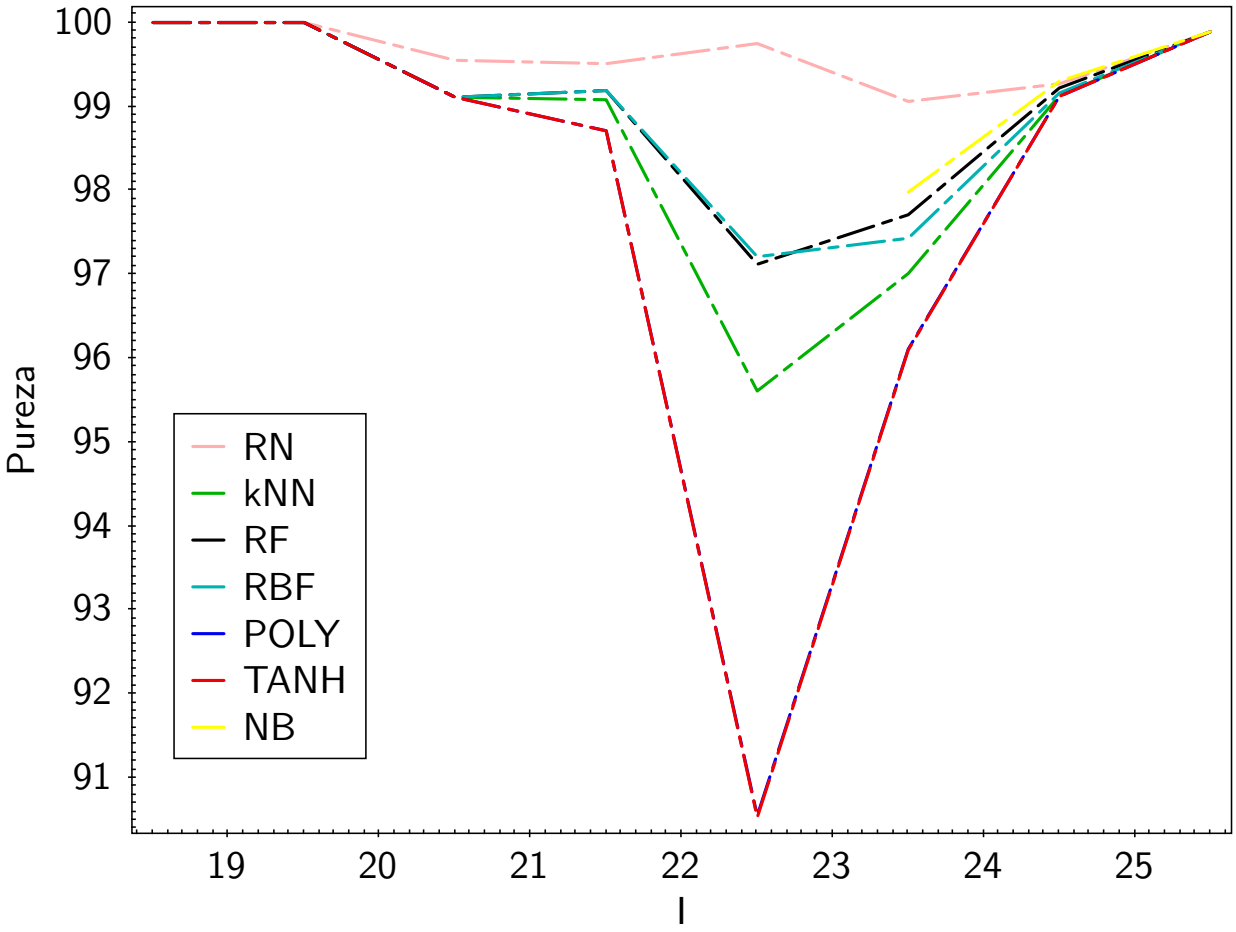

Figura III.5: Magnitude *i* vs Pureza

A figura III.6 apresenta as faixas de magnitude pela completeza de cada classificador. O comportamento do SVM tanh é anomalo, pois coincide com o do ZeroR. Para os demais classificadores, todos também exibem uma queda de desempenho entre as magnitudes 22 e 24, provavelmente pelos mesmos efeitos descritos no parágrafo anterior.

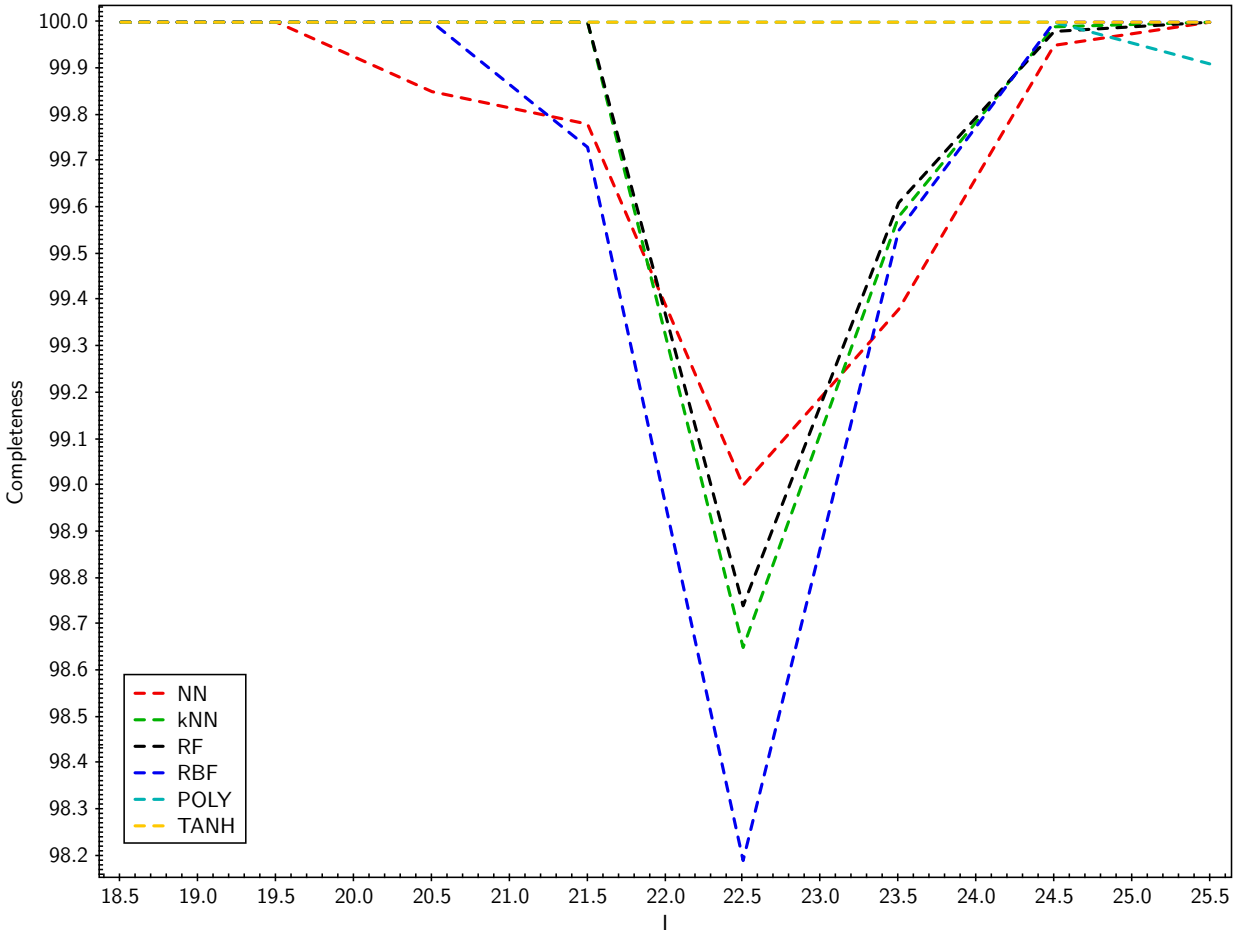

Figura III.6: Magnitude *i* vs Completeza

Capítulo IV Conclusões

Este capítulo é composto das seções de Retrospectiva (seção IV.1), Contribuição (seção IV.2), Legado (seção IV.3) e Trabalhos Futuros (seção IV.4).

IV.1 Retrospectiva

O presente trabalho propõe empregar diversos métodos de aprendizado de máquina tais como NN, SVM, kNN, RF e NB na tentativa de obter um desempenho otimizado no problema da separação estrela/galáxia.

De acordo com o material consultado para servir de referência ao trabalho (artigos científicos principalmente), verificou-se que poucos autores utilizaram, ou pelo menos reportaram, técnicas de pré-processamento de dados antes da geração de um modelo de classificação. Esta lacuna foi observada e serviu de motivação ao enfoque diferencial dado neste trabalho. Sendo assim, definiu-se um *workflow* onde as primeiras atividades seriam de pré-processamento, a saber, limpeza de dados, amostragem, remoção de outliers, normalização e criação de folds para o uso com validação cruzada.

A geração do modelo com validação cruzada permitiu estimar, para cada configuração particular, de cada método, um valor otimizado de desempenho (no caso a área ROC). A escolha de parâmetros foi analisada e cada método teve um número de parâmetros para configurar que variou de zero (NB) até três (números de neurônios, decaimento e número máximo de iterações) no caso de NN. O número de configurações possíveis para cada método foi igualado para permitir um comparação justa entre eles.

Durante o estudo foi possível observar o comportamento dos diferentes métodos de aprendizado de máquina sob a ótica da otimização de hiperparâmetros, técnica pouco reportada na literatura científica da área. Isso demonstrou a importância desse refinamento no alcance do resultado final, especialmente no caso do kNN e RF, altamente sensíveis aos parâmetros de entrada (número de vizinhos e número de árvores respectivamente). Ficou claro que a otimização contribui de modo decisivo para a melhora do desempenho global. A desvantagem dessa abordagem está no tempo de processamento necessário para a otimização, em especial no caso do método SVM. Após a otimização, o modelo de classificação gerado utilizando a melhor configuração de

parâmetros para cada método foi então aplicada ao dataset de teste para a avaliação comparativa dos métodos.

IV.2 Contribuição

Este trabalho contribui por abordar o problema de prover uma análise comparativa de vários métodos de aprendizado de máquina para a separação estrela/galáxia baseada em análise fotométrica de catálogos astronômicos. Os experimentos revelaram que em termos de acurácia, a maioria dos métodos analisados e explorados tiveram desempenho superior ao método *baseline* ZeroR. Entre eles, tanto NN como RF alcançaram um bom desempenho. Este também foi o caso quando a métrica empregada foi a área ROC. Para a completeza e pureza esses métodos também apresentaram um bom balanceamento de resultados.

IV.3 Legado

Na preparação deste trabalho muitas lições foram aprendidas. Na parte técnica, o contato com uma nova linguagem de programação - a linguagem R. O uso desta linguagem para a geração de código foi uma escolha natural devido a sua flexibilidade e pacotes disponíveis. Além disso, a compreensão de como fazer um *workflow* foi adquirida. Na parte teórica, foram estudados vários aspectos de Mineração de Dados que complementaram a formação obtida na graduação. Concomitantemente, dois artigos científicos foram escritos sobre o tema abordado, um para o congresso da SBC e outro para a IJCNN, o que exigiu disciplina e maturidade para sua elaboração. Durante a pesquisa de trabalhos relacionados, ficou evidente que lacunas pertinentes ao tema existiam e que deviam ser exploradas, o que em parte foi feito.

IV.4 Trabalhos Futuros

Como trabalhos futuros, planeja-se utilizar a transformação de atributos do dataset COSMOS usando Principal Components Analysis (PCA) antes das atividades do *workflow*. O uso do PCA é uma alternativa a seleção de atributos usada neste trabalho. PCA quando exposta a um grande conjunto de variáveis correlacionadas permite sumarizar este conjunto com um número menor de variáveis representativas que coletivamente explicam a maioria da variabilidade do conjunto inicial. A direção dos componentes principais são direções no espaço característico com o qual os dados originais são altamente variáveis. PCA então se refere ao processo pelo qual os componentes principais são computados e o subsequente uso desses componentes para entender os dados [Lantz, 2013].

A seleção de atributos feita por métodos de aprendizado estatístico tais como Forward Stepwise

Selection e Lasso também pode ser considerada. Forward Stepwise Selection começa com um modelos sem preditores (atributos) e então adiciona preditores ao modelo, um por vez, até que todos os preditores estejam no modelo. Em particular, a cada passo a variável que dá o melhor melhoramento adicional ao ajuste é adicionada ao modelo. O melhor modelo é identificado entre o conjunto de modelos com diferentes números de variáveis.

Em contraste com esse método, pode-se ajustar um modelo contendo todos os p preditores usando uma técnica que restringe ou regulariza a estimativa dos coeficientes, ou equivalentemente, que encolhe a estimativa dos coeficientes para zero, como é o caso do Lasso [Lantz, 2013]. Também se pretende investigar o uso de métodos *ensemble* [Seni and IV, 2010], que leva em conta a decisão de mais de um classificador para escolher a classe correta.

Referências Bibliográficas

Andreon, S., Gargiulo, G., Longo, G., Tagliaferri, R., and Capuano, N. (2000). Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification. , 319:700–716.

Andreon, S., Gargiulo, G., Longo, G., Tagliaferri, R., and Capuano, N. (2002). Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification: Wide field imaging - I. Monthly Notices of the Royal Astronomical Society, 319(3):700–716.

Ball, N. M., Brunner, R. J., Myers, A. D., and Tcheng, D. (2006). Robust machine learning applied to astronomical data sets. I. star-galaxy classification of the Sloan Digital Sky Survey DR3 using decision trees. The Astrophysical Journal, 650(1):497.

Bazell, D. and Peng, Y. (1998). A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination. The Astrophysical Journal Supplement Series, 116(1):47–55.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems, pages 2546–2554.

Bhatia, N. and others (2010). Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085.

Chirigati, F., Silva, V., Ogasawara, E., de Oliveira, D., Dias, J., Porto, F., Valduriez, P., and Mattoso, M. (2012). Evaluating parameter sweep workflows in high performance computing. In Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies, page 2. ACM.

Cortiglioni, F., Mahonen, P., Hakala, P., and Frantti, T. (2001). Automated Star-Galaxy Discrimination for Large Surveys. The Astrophysical Journal, 556(2):937–943.

Dawson, K. S., Kneib, J.-P., Percival, W. J., Alam, S., Albareti, F. D., Anderson, S. F., Armengaud, E., Aubourg, É., Bailey, S., Bautista, J. E., Berlind, A. A., Bershady, M. A., Beutler, F., Bizyaev, D., Blanton, M. R., Blomqvist, M., Bolton, A. S., Bovy, J., Brandt, W. N., Brinkmann,

J., Brownstein, J. R., Burtin, E., Busca, N. G., Cai, Z., Chuang, C.-H., Clerc, N., Comparat, J., Cope, F., Croft, R. A. C., Cruz-Gonzalez, I., da Costa, L. N., Cousinou, M.-C., Darling, J., de la Macorra, A., de la Torre, S., Delubac, T., du Mas des Bourboux, H., Dwelly, T., Ealet, A., Eisenstein, D. J., Eracleous, M., Escoffier, S., Fan, X., Finoguenov, A., Font-Ribera, A., Frinchaboy, P., Gaulme, P., Georgakakis, A., Green, P., Guo, H., Guy, J., Ho, S., Holder, D., Huehnerhoff, J., Hutchinson, T., Jing, Y., Jullo, E., Kamble, V., Kinemuchi, K., Kirkby, D., Kitaura, F.-S., Klaene, M. A., Laher, R. R., Lang, D., Laurent, P., Le Goff, J.-M., Li, C., Liang, Y., Lima, M., Lin, Q., Lin, W., Lin, Y.-T., Long, D. C., Lundgren, B., MacDonald, N., Geimba Maia, M. A., Malanushenko, E., Malanushenko, V., Mariappan, V., McBride, C. K., McGreer, I. D., Ménard, B., Merloni, A., Meza, A., Montero-Dorta, A. D., Muna, D., Myers, A. D., Nandra, K., Naugle, T., Newman, J. A., Noterdaeme, P., Nugent, P., Ogando, R., Olmstead, M. D., Oravetz, A., Oravetz, D. J., Padmanabhan, N., Palanque-Delabrouille, N., Pan, K., Parejko, J. K., Pâris, I., Peacock, J. A., Petitjean, P., Pieri, M. M., Pisani, A., Prada, F., Prakash, A., Raichoor, A., Reid, B., Rich, J., Ridl, J., Rodriguez-Torres, S., Carnero Rosell, A., Ross, A. J., Rossi, G., Ruan, J., Salvato, M., Sayres, C., Schneider, D. P., Schlegel, D. J., Seljak, U., Seo, H.-J., Sesar, B., Shandera, S., Shu, Y., Slosar, A., Sobreira, F., Streblyanska, A., Suzuki, N., Taylor, D., Tao, C., Tinker, J. L., Tojeiro, R., Vargas-Magaña, M., Wang, Y., Weaver, B. A., Weinberg, D. H., White, M., Wood-Vasey, W. M., Yeche, C., Zhai, Z., Zhao, C., Zhao, G.-b., Zheng, Z., Ben Zhu, G., and Zou, H. (2016). The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. , 151:44.

de Fátima Oliveira Saraiva, M. (2004). Astronomia & Astrofísica. LIVRARIA DA FISICA.

Djorgovski, S., Donalek, C., Mahabal, A., Williams, R., Drake, A., Graham, M., and Glikman, E. (2006). Some Pattern Recognition Challenges in Data-Intensive Astronomy. pages 856–863. IEEE.

Elting, C., Bailer-Jones, C. A. L., Smith, K. W., and Bailer-Jones, C. A. (2008). Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines. pages 9–14. AIP.

Fadely, R., Hogg, D. W., and Willman, B. (2012). Star-Galaxy Classification in Multi-Band Optical Imaging. The Astrophysical Journal, 760(1):15.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8):861–874.

Flaugher, B., Diehl, H. T., Honscheid, K., Abbott, T. M. C., Alvarez, O., Angstadt, R., Annis, J. T., Antonik, M., Ballester, O., Beaufore, L., Bernstein, G. M., Bernstein, R. A., Bigelow, B., Bonati, M., Boprie, D., Brooks, D., Buckley-Geer, E. J., Campa, J., Cardiel-Sas, L., Castander, F. J.,

Castilla, J., Cease, H., Cela-Ruiz, J. M., Chappa, S., Chi, E., Cooper, C., da Costa, L. N., Dede, E., Derylo, G., DePoy, D. L., de Vicente, J., Doel, P., Drlica-Wagner, A., Eiting, J., Elliott, A. E., Emes, J., Estrada, J., Fausti Neto, A., Finley, D. A., Flores, R., Frieman, J., Gerdes, D., Gladders, M. D., Gregory, B., Gutierrez, G. R., Hao, J., Holland, S. E., Holm, S., Huffman, D., Jackson, C., James, D. J., Jonas, M., Karcher, A., Karliner, I., Kent, S., Kessler, R., Kozlovsky, M., Kron, R. G., Kubik, D., Kuehn, K., Kuhlmann, S., Kuk, K., Lahav, O., Lathrop, A., Lee, J., Levi, M. E., Lewis, P., Li, T. S., Mandrichenko, I., Marshall, J. L., Martinez, G., Merritt, K. W., Miquel, R., Muñoz, F., Neilsen, E. H., Nichol, R. C., Nord, B., Ogando, R., Olsen, J., Palaio, N., Patton, K., Peoples, J., Plazas, A. A., Rauch, J., Reil, K., Rheault, J.-P., Roe, N. A., Rogers, H., Roodman, A., Sanchez, E., Scarpine, V., Schindler, R. H., Schmidt, R., Schmitt, R., Schubnell, M., Schultz, K., Schurter, P., Scott, L., Serrano, S., Shaw, T. M., Smith, R. C., Soares-Santos, M., Stefanik, A., Stuermer, W., Suchyta, E., Sypniewski, A., Tarle, G., Thaler, J., Tighe, R., Tran, C., Tucker, D., Walker, A. R., Wang, G., Watson, M., Weaverdyck, C., Wester, W., Woods, R., Yanny, B., and The DES Collaboration (2015). The Dark Energy Camera. \aj, 150:150.

Fukugita, M., Ichikawa, T., Gunn, J., Doi, M., Shimasaku, K., and Schneider, D. (1996). The sloan digital sky survey photometric system. The Astronomical Journal, 111:1748.

Gao, D., Zhang, Y.-X., and Zhao, Y.-H. (2009). Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics, 9(2):220.

Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann, Waltham, Mass., 3 edition edition.

Henden, A. A. and Kaitchuck, R. H. (1982). Astronomical photometry. New York, Van Nostrand Reinhold Co., 1982. 405 p., 1.

Hubble, E. (1929). A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. Proceedings of the National Academy of Science, 15:168–173.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer, 1st ed. 2013. corr. 4th printing 2014 edition edition.

Jiang, L., Wang, D., Cai, Z., and Yan, X. (2007). Survey of improving naive Bayes for classification. In Advanced Data Mining and Applications, pages 134–145. Springer.

Kim, E. J., Brunner, R. J., and Kind, M. C. (2015). A Hybrid Ensemble Learning Approach to Star-Galaxy Classification. arXiv preprint arXiv:1505.02200.

Kovács, A. and Szapudi, I. (2013). Star-galaxy separation strategies for WISE-2mass all-sky infrared galaxy catalogs. arXiv:1401.0156 [astro-ph]. arXiv: 1401.0156.

Lantz, B. (2013). Machine Learning with R. Packt Publishing, Birmingham.

Li, L., Zhang, Y., and Zhao, Y. (2008). k-Nearest Neighbors for automated classification of celestial objects. Science in China Series G: Physics, Mechanics and Astronomy, 51(7):916–922.

Mahonen, P. and Frantti, T. (2000). Fuzzy Classifier for Star?Galaxy Separation. The Astrophysical Journal, 541(1):261–263.

Odewahn, S. and Nielsen, M. (1994). Star-galaxy separation using neural networks. Vistas in Astronomy, 38:281–IN4.

Odewahn, S., Stockwell, E., Pennington, R., Humphreys, R., and Zumach, W. (1992). Automated star/galaxy discrimination with neural networks. In Digitised Optical Sky Surveys, pages 215–224. Springer.

Ogasawara, E., Murta, L., Zimbrao, G., and Mattoso, M. (2009). Neural networks cartridges for data mining on time series. In International Joint Conference on Neural Networks, 2009. IJCNN 2009, pages 2302–2309.

O'Keefe, P. J., Gowanlock, M. G., McConnell, S. M., and Patton, D. (2009). Star-Galaxy Classification Using Data Mining Techniques with Considerations for Unbalanced Datasets. In Astronomical Data Analysis Software and Systems XVIII, volume 411, page 318.

Philip, N. S., Wadadekar, Y., Kembhavi, A., and Joseph, K. B. (2002). A difference boosting neural network for automated star-galaxy classification. Astronomy and Astrophysics, 385(3):1119–1126.

Schlegel, D. J., Finkbeiner, D. P., and Davis, M. (1998). Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds. , 500:525–553.

Scoville, N., Aussel, H., Brusa, M., Capak, P., Carollo, C., Elvis, M., Giavalisco, M., Guzzo, L., Hasinger, G., Impey, C., and others (2007). The cosmic evolution survey (COSMOS): overview. The Astrophysical Journal Supplement Series, 172(1):1.

Seni, G. and IV, J. F. E. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

Smith, J. A., Tucker, D. L., Kent, S., Richmond, M. W., Fukugita, M., Ichikawa, T., Ichikawa, S.-i., Jorgensen, A. M., Uomoto, A., Gunn, J. E., and others (2002). The ugriz standard-star system. The Astronomical Journal, 123(4):2121.

Soumagnac, M. T., Abdalla, F. B., Lahav, O., Kirk, D., Sevilla, I., Bertin, E., Rowe, B. T. P., Annis, J., Busha, M. T., Da Costa, L. N., Frieman, J. A., Gaztanaga, E., Jarvis, M., Lin, H., Percival, W. J., Santiago, B. X., Sabiu, C. G., Wechsler, R. H., Wolz, L., and Yanny, B. (2015). Star/galaxy separation at faint magnitudes: application to a simulated Dark Energy Survey. \mnras, 450:666–680.

Vasconcellos, E. C., Carvalho, R. R. d., Gal, R. R., LaBarbera, F. L., Capelato, H. V., Velho, H. F. C., Trevisan, M., and Ruiz, R. S. R. (2011). Decision Tree Classifiers for Star/Galaxy Separation. The Astronomical Journal, 141(6):189.

Verikas, A., Gelzinis, A., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. Pattern Recognition, 44(2):330–349.

Wang, J. (2009). Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval. In Boughanem, M., Berrut, C., Moth, J., and Soulé-Dupuy, C., editors, ECIR, volume 5478 of Lecture Notes in Computer Science, pages 4–16. Springer.

Zhang, Y. and Zhao, Y. (2003). Classification in multidimensional parameter space: Methods and examples. Publications of the Astronomical Society of the Pacific, 115(810):1006–1018.

Zhao, Y. and Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12):1955–1959.