Classificação de Objetos Astronômicos em Estrelas e Galáxias Usando o Algoritmo J4.8

Renata S. Rocha[1], Haroldo F. Campos Velho[1], Rafael D. C. dos Santos[1], Marina Trevisan[2]

[1]Laboratório Associado de Computação e Matemática Aplicada – Instituto Nacional de Pesquisas Espaciais (INPE)
12227-010 – São José dos Campos – SP – BRASIL

[2]Departamento de Astronomia, IAG, USP
05508-090, São Paulo, SP

{renata,haroldo,rafael.santos}@lac.inpe.br, trevisan@astro.iag.usp.br

Abstract. *The optical measurement data constitute a source of very important information for the astronomy. Such measurements is fundamental to classify stars and galaxies. This work describes the algorithm to design decision trees (J4.8 algorithm). The classifiers were employed to the astronomical data from the project Sloan Digital Sky Survey (SDSS). The performance for the best classifiers for the test set was greater than 98% for stars classification, and greater than 99% for galaxies classification.*

Resumo. *Os registros de astronomia ótica constituem uma fonte de informação extremamente importante. Estas medidas são fundamentais para classificar estrelas e galáxias. Este trabalho descreve o algoritmo de construção de árvore de decisão (J4.8). Dados do projeto Sloan Digital Sky Survey (SDSS) foram usados para treinamento e validação dos classificadores desenvolvidos. Os classificadores apresentaram índices de acerto, sobre o conjunto de teste, superiores a 98% para a classificação de estrelas e superiores a 99% para a classificação de galáxias.*

1. Introdução

O entendimento sobre a origem e evolução do Universo tem sido alterado ao longo dos tempos. Na cultura ocidental, antes da era moderna, o ponto de vista aristotélico prevalecia, ou seja, existia uma física para os fenômenos da Terra e outra física para os corpos celestiais. O inglês Isaac Newton alterou para sempre este paradigma e desenvolveu um modelo físico-matemático constituído de poucos postulados e algumas leis, entre estas leis a formulação matemática da gravitação Universal. O modelo cosmológico de Newton é de um Universo infinito, se assim não fosse, o sistema todo iria colapsar sobre um centro, o que contraria a observação de um Universo aparentemente estável e estático.

No início do século passado, em 1917, Albert Einstein propôs um modelo cosmológico relativístico, ainda considerando o Universo como estático. Alguns anos depois, dados de observação aliados a teoria permitiram uma das mais importantes descobertas da astronomia no século XX: o Universo está em expansão. Esta descoberta foi realizada por Hubble em 1929, quando ele observando uma classe de estrelas conhecidas como *Cepheids* descobriu que as galáxias estão se afastando. Tal descoberta marca o fim da

era de um Universo estático e o início de uma nova era. Surge, então, o modelo cosmológico de um Universo em expansão [Carvalho et al. 2008, Madsen 1996].

Conforme Madsen (1996), a dinâmica composta pela força da gravidade e pela expansão do Universo descreve a história da formação das grandes estruturas cosmológicas (galáxias, aglomerados de galáxias, super-aglomerados, etc.). As grandes bases de dados astronômicos existentes hoje fornecem uma possibilidade de estudo dessas estruturas sem precedentes. Porém, o estudo dessas estruturas depende do correto mapeamento de galáxias, mas numa imagem astronômica, nem sempre é fácil fazer a distinção entre uma galáxia e uma estrela. A dificuldade é porque quanto mais distante está uma galáxia do nosso planeta, menor é o seu tamanho na imagem e menor é a luminosidade observada. Quando se atinge um limite crítico de tamanho e luminosidade é difícil distinguir entre uma galáxia muito distante e uma estrela de baixa luminosidade da nossa própria galáxia [Carvalho et al. 2008]. A Figura 1 ilustra a dificuldade mencionada, nesta figura pode-se observar que na parte superior é simples realizar a identificação de um objeto de alta luminosidade, na parte central a identificação ainda é simples, mas o número de objetos nesse domínio de luminosidade é muito grande para ser feito visualmente, na parte inferior a identificação é muito complexa e necessita de métodos sofisticados. Tais métodos se baseiam em um conjunto de parâmetros que descrevem a imagem. Esses parâmetros podem ser fotométricos ou espectroscópicos. Porém, a aquisição de espectros em geral requer um tempo maior de observação e utilizar dados fotométricos tornam as observações mais eficientes.

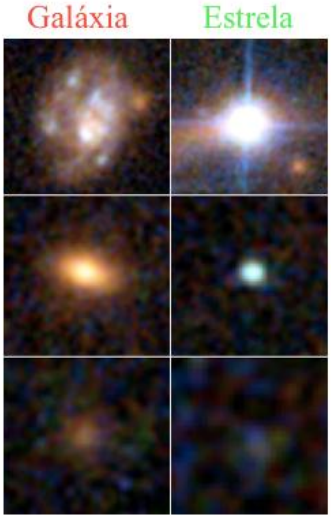

Figura 1. Separação estrela-galáxia:
Fonte: Carvalho et al. (2008).

Neste contexto, separar estrelas de galáxias a partir de dados fotométricos é um desafio bastante interessante e o objetivo deste trabalho é aplicar a técnica de árvores de decisão a este problema. De um modo geral, existe na literatura uma série de trabalhos utilizando árvores de decisão para classificar objetos astronômicos [Bazell e Aha 2001, Salzberg et al. 1995, Zhang e Zhao 2007, Zhao e Zhang 2008].

Em particular, para dados do projeto *Sloan Digital Sky Survey* (SDSS) uma abordagem em árvores de decisão foi utilizada por Suchkov et al. (2005) na classificação de objetos fotométricos em estrelas, galáxias e Núcleos Galácticos Ativos (AGN). Para o desenvolvimento do modelo foi utilizado o projeto ClassX. Este projeto é um sistema

online que foi originalmente desenvolvido para realizar a classificação de fontes de raio X. O algoritmo de criação da árvore de decisão utilizado pelo ClassX é o sistema OC1 de Murthy et al. (1994).

Ball et al. (2006) utilizaram árvores de decisão para realizar a classificação de objetos da terceira divulgação de dados do SDSS em estrelas, galáxias ou "nem estrela nem galáxia". O classificador foi treinado sobre um conjunto de 477.068 objetos espectroscopicamente classificados. Posteriormente o classificador desenvolvido foi utilizado para classificação de cerca de 143 milhões de objetos do projeto SDSS, sendo este o primeiro trabalho a utilizar árvores de decisão para classificar um conjunto inteiro de dados do SDSS.

O objetivo deste trabalho é utilizar o algoritmo de construção de árvore de decisão C4.5, implementado no *software Waikato Environment for Knowledge Analysis* (WEKA) como J4.8 [Witten e Frank 2000], para desenvolver classificadores baseados em atributos fotométricos para serem utilizados na classificação de objetos astronômicos do projeto SDSS em estrelas ou galáxias.

As demais seções deste trabalho estão divididas da seguinte maneira: Na Seção 2 tem-se a descrição do algoritmo C4.5, a Seção 3 apresenta os dados utilizados no treinamento e validação dos classificadores, os resultados obtidos são apresentados na Seção 4. Finalmente a Seção 5 é reservada as considerações finais.

2. Algoritmo de Indução C4.5 ou J4.8

Considere um conjunto de objetos que são descritos em termos de uma coleção de atributos. Estes objetos podem pertencer a diferentes classes, cada atributo mede alguma característica importante de um objeto. Considere também um conjunto de treinamento, cuja classe de cada objeto é conhecida. Conforme Quinlan (1986) é possível desenvolver uma regra de classificação que pode determinar a classe de qualquer objeto a partir dos valores dos seus atributos. Tal regra de classificação pode ser expressa como uma árvore de decisão. Uma árvore de decisão é uma estrutura simples em que as folhas contêm as classes, os nós não-folhas representam atributos baseados em testes com um ramo para cada possível saída [Hunt et al. 1966, Quinlan 1986, Quinlan 1993]. Para classificar um objeto, começa-se com a raiz da árvore, aplica-se o teste e toma-se o ramo apropriado para aquela saída. O processo continua e quando uma folha é encontrada o objeto é classificado segundo a classe indicada naquela folha.

Como formar uma árvore para uma coleção arbitrária de C objetos? Se C é vazio ou contém somente objetos de uma classe, a árvore de decisão mais simples contém uma folha que representa essa classe. Caso contrário, seja T algum teste sobre um objeto que tem os possíveis resultados $O_1, O_2, ..., O_W$. Cada objeto em C dá um desses resultados para T, portanto T produz uma partição $\{C_1, C_2, ..., C_W\}$ de C, com C_i contendo os objetos que tem saída O_i. No pior caso essa estratégia fornecerá subconjuntos de um único objeto.

No caso de uma amostra de objetos que pertencem somente a duas classes, por exemplo, P e N, um objeto qualquer pertencerá a classe P com probabilidade $p/(p+n)$ e a classe N com probabilidade $n/(p+n)$, em que p é o número total de objetos pertencentes a classe P e n é número total de objetos pertencentes a classe N. Quando uma árvore de decisão é usada para classificar um objeto, ela retorna uma classe. Assim, ela pode ser

considerada como uma fonte de mensagem *P* ou *N* em que a informação necessária para gerar a mensagem é obtida conforme a Equação 2.1:

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \qquad (2.1)$$

Se o atributo A com os valores $[A_1, A_2, \ldots, A_V]$ é usado para a raiz da árvore de decisão, ele dividirá *C* em $[C_1, C_2, \ldots, C_V]$, onde C_i contém os objetos em *C* que têm valores A_i de *A*. Considere C_i contendo p_i objetos da classe *P* e n_i objetos da classe *N*, a informação necessária para a sub-árvore em C_i é $I(p_i, n_i)$. A informação necessária para a árvore com *A* como raiz é obtida com a média ponderada:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i) \qquad (2.2)$$

O ganho de informação obtido pelo ramo usando o atributo *A* é dado pela Equação 2.3.

$$G(A) = I(p,n) - E(A) \qquad (2.3)$$

O algoritmo C4.5 examina todos os atributos candidatos e escolhe *A* que maximiza o ganho de informação. O processo é repetido recursivamente para obter os demais nós e formar a árvore de decisão com os subconjuntos restantes [Quinlan 1993].

3. Aquisição dos Dados e parâmetros Utilizados

Os dados utilizados neste trabalho são dados de seis anos do projeto SDSS [Adelman – McCarthy et al. 2008]. Este levantamento cobre uma área de aproximadamente 10.000 graus quadrados do céu, contendo imagens de 287 milhões de objetos. A câmera do SDSS mede quão brilhantes são os objetos em cinco bandas fotométricas denominadas *u*, *g*, *r*, *i*, *z*. Além disso, há também o levantamento espectroscópico, que cobre uma área de aproximadamente 7.500 graus quadrados, com mais de um milhão de objetos catalogados.

A classificação de objetos baseada nos espectros é mais confiável. Sendo assim, os objetos do catálogo fotométrico foram selecionados levando em conta também as informações do catálogo espectroscópico, de forma a minimizar a falsa classificação de objetos nas amostras de treinamento e de testes. Cada objeto identificado nas imagens é classificado pelo próprio *pipeline*[1] do SDSS, em todas as cinco bandas. O mesmo é feito para os dados espectroscópicos. Baseando-se nisso, o critério de seleção foi exigir que o a classificação do objeto seja a mesma usando os algoritmos do *pipeline* e a obtida espectroscopicamente. A amostra final, atendendo essa restrição, é composta por 43.289 estrelas e 452.400 galáxias. Os dados foram obtidos através do servidor *CasJobs* do SDSS [Lin e Thakar 2008, Szalay et al. 2008, Thakar et al. 2008]. Dos objetos da amostra final foram selecionados os parâmetros fotométricos considerados relevantes na distinção entre estrelas e galáxias, em todas as cinco bandas. A seguir uma breve descrição desses parâmetros:

---

[1] *Pipeline*: Esta expressão denota uma série de procedimentos automáticos: aquisição dos dados, processamento da imagem (remoção de ruídos), extração de atributos (magnitudes, calibração de fluxos fotométricos, entre outros) e classificação de objetos.

- **Nprof:** De cada objeto é extraído o perfil radial de brilho superficial. Este perfil é dado como a média azimutal do brilho em uma série de anéis, cujos raios podem ser encontrados na tabela 7 de [Stoughton et al. 2002]. O parâmetro *nprof* corresponde ao número de anéis para os quais ainda existe um sinal mensurável.
- **PetroR50** e **PetroR90:** Para cada objeto é definido o perfil de brilho superficial Petrosiano [Petrosian 1976], a partir deste são definidos os raios *PetroR50* e *PetroR90*, que correspondem aos raios que compreendem 50% e 90% do fluxo Petrosiano, respectivamente. De uma maneira simplificada, estes podem ser entendidos como uma medida da "extensão" do objeto. Objetos mais difusos como galáxias tendem a ter o raio petrosiano maior.
- **isoA, isoB:** Os atributos isoA e isoB são definidos como o eixo maior e o eixo menor da figura geométrica representativa do objeto e são utilizados para encontrar a excentricidade. Logo, ambos se convertem em um único parâmetro a ser utilizado no treinamento.
- **Magnitudes:** Foram utilizadas as magnitudes Petromag, PSFmag, Fibermag, Modelmag. Magnitude é uma medida do brilho aparente do objeto e cada uma das quatro magnitudes são obtidas considerando modelos diferentes para o perfil de brilho: perfil petrosiano, perfil da *Point Spread Function*, da fibra ótica (dado espectroscópico) e a magnitude baseada no modelo que melhor se ajusta. Uma descrição detalhada das magnitudes pode ser encontrada em [Stoughton et al. 2002].

Assim, cada objeto do conjunto de treinamento é descrito em termos de 40 parâmetros (8 atributos x 5 bandas).

4. Resultados Obtidos

O treinamento (criação da árvore) foi realizado com um conjunto de 10.000 objetos (925 estrelas e 9.075 galáxias). Utilizando-se o algoritmo J4.8 várias árvores foram criadas, analisadas e modificadas (variando-se o número mínimo de objetos por folha, fator de confiança usado na poda, entre outros) para evitar que regras muito complexas e que correspondem a poucos casos no conjunto de dados fossem criadas. Nesse trabalho, será apresentado o desempenho de duas árvores testadas sobre o conjunto total de amostras (495. 689 objetos). Tais árvores foram implementadas em linguagem C, já que a interface WEKA possui limitação de memória, não sendo possível sua utilização em um conjunto grande de dados. As Figuras 2 e 3 exibem as árvores criadas.

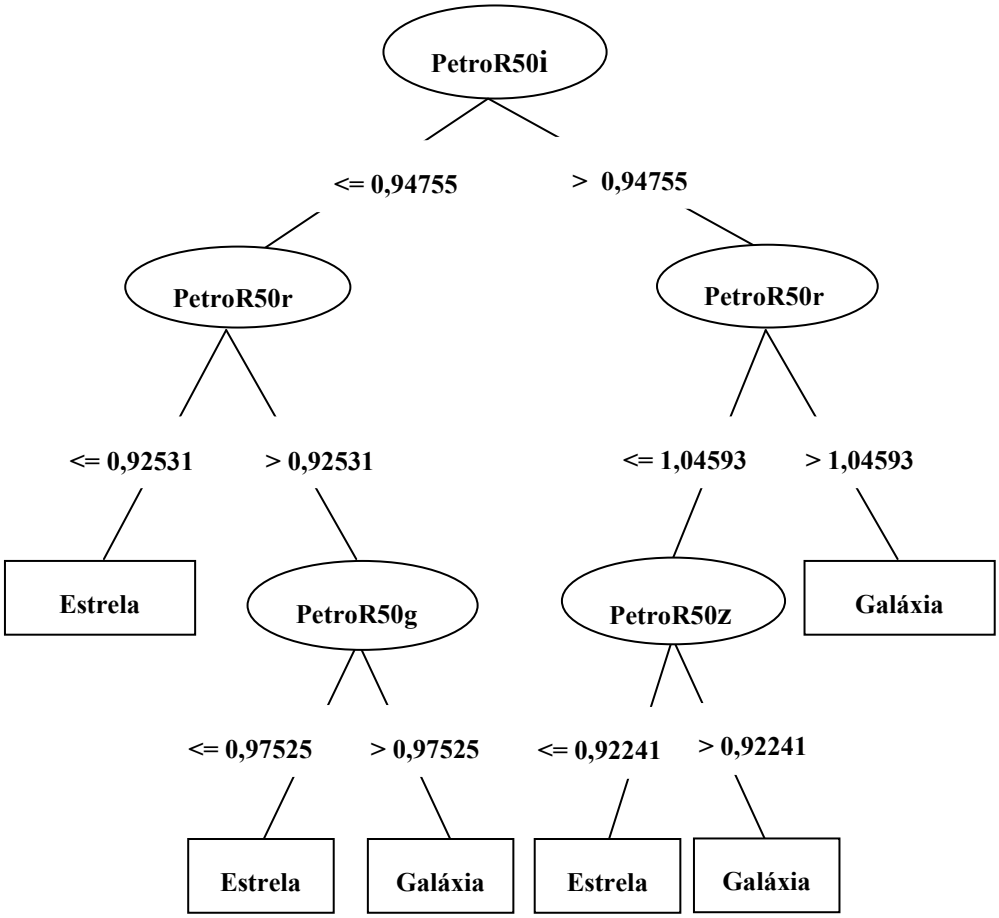

Figura 2. Primeira árvore criada com o algoritmo J4.8 para classificação de objetos astronômicos em estrelas e galáxias e com no mínimo 5 objetos por folha.

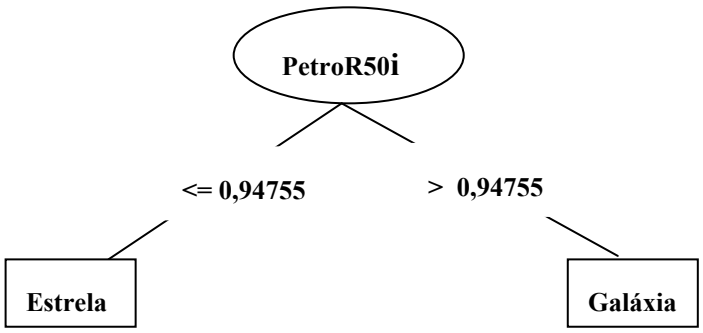

Figura 3. Segunda árvore criada com o algoritmo J4.8 para classificação de objetos astronômicos em estrelas e galáxias e com no mínimo 20 objetos por folha

Estas árvores indicam que o atributo fotométrico ***PetroR50*** nas bandas ***i*** e ***r*** permite uma melhor separação entre as classes, ou seja, contêm a maior quantidade de informação. A Tabela 1 apresenta os resultados da classificação das duas árvores sobre o conjunto de teste, pode-se observar que os resultados usando a primeira árvore indicam que 523 estrelas do total de 43.289 foram classificadas erroneamente como galáxias e 24 estrelas não foram classificadas devido a ausência de algum atributo necessário a sua classificação. Na classificação de galáxias, os resultados mostram que 1.866 galáxias do

total de 452.400 foram classificadas erroneamente como estrelas e 52 galáxias não foram classificadas, também devido a ausência de algum atributo. O índice de acerto para a classificação de estrelas em termos de porcentagem foi de 98,79% e para a classificação de galáxias foi de 99,59%. Os resultados obtidos com a segunda árvore mostram que 568 estrelas foram classificadas erroneamente como galáxias e 13 estrelas não foram classificadas. Na classificação de galáxias, 2.344 foram classificadas erroneamente como estrelas e 45 não foram classificadas. O índice de acerto para a classificação de estrelas foi de 98,69% e para a classificação de galáxias foi de 99,48%.

Tabela 1. Resultados obtidos para as duas árvores com o conjunto de 495.689 objetos.

	1ª árvore		2ª árvore	
	Estrelas	**Galáxias**	**Estrelas**	**Galáxias**
Estrelas	42.742	523	42.708	568
Galáxias	1.866	450.482	2.344	450.011
Objetos não classificados	24	52	13	45
Índice de acerto	98,79%	99,59%	98,69%	99,48%

Pode ser observado na Tabela 2 o índice de acertos referentes a classificação de estrelas e galáxias do projeto SDSS obtidos com os classificadores desenvolvidos neste trabalho e também com os trabalhos de Suchkov et al. (2005) e Ball et al. (2006).

Os parâmetros fotométricos utilizados em ambos os trabalhos da literatura foram as cores dos objetos. A cor de um objeto pode ser medida através das diferenças de magnitudes entre os filtros. Suchkov et al. (2005) usaram as diferenças $u-g$, $g-r$, $r-i$, $i-z$ e $g-i$ e Ball et al. (2006) utilizaram as mesmas cores que Suchkov et al. (2005) com exceção de $g-i$. Conforme pode ser observado, os classificadores baseados em árvores de decisão mencionados neste trabalho utilizando parâmetros fotométricos, apresentaram um desempenho compatível ao obtido nos trabalhos de Suchkov et al. (2005) e Ball et al. (2006). O índice de acerto na classificação de estrelas foi cerca de 0,60% superior ao trabalho de Suchkov et al. (2005) e cerca de 5, 40% superior ao trabalho de Ball et al. (2006). Para a classificação de galáxias o índice de acerto foi cerca de 1,00% superior a ambos os trabalhos.

Tabela 2. Comparação dos resultados obtidos neste trabalho com os trabalhos de Suchkov et al. (2005) e Ball et al. (2006).

Classificadores	Parâmetros utilizados	Índice de acerto	
		Estrelas	Galáxias
Suchkov et. al (2005)	Cores dos objetos	98,10%	98,50%
Ball et. al (2006)	Cores dos objetos	93,40%	98,20%
1ª árvore	Raio ***PetroR50*** (bandas i, r, g e z)	98,79%	99,59%
2ª árvore	Raio ***PetroR50*** (banda i)	98,69%	99,48%

5. Considerações Finais

Resultados preliminares com a técnica de árvores de decisão foram obtidos na classificação de estrelas e galáxias para dados do projeto SDSS, com base em parâmetros fotométricos, onde o algoritmo de construção empregado foi o C4.5 implementado no *software* WEKA como J4.8. De um modo geral, a estratégia do sistema de árvores de decisão é um sistema automático de projeto de um classificador, baseado em aprendizado de máquina, que tornam explícitos os atributos mais relevantes.

Existem algumas diferenças entre os índices de acerto obtidos com o presente trabalho e os resultados da literatura. Estas diferenças podem ser atribuídas a vários fatores, entre eles: (a) nos trabalhos da literatura outras classes foram consideradas (e não somente estrela/galáxia); (b) há diferenças nos parâmetros fotométricos analisados e/ou limiares considerados; (c) estratégias da configuração de árvores de decisão (número mínimo de objetos por folhas, dentre outros).

Os classificadores desenvolvidos com árvores de decisão no presente trabalho alcançaram desempenho similar aos classificadores desenvolvidos por Suchkov et al. (2005) e Ball et al. (2006). Esses resultados mostram que o algoritmo de indução testado é robusto para o desenvolvimento de classificadores com base em atributos fotométricos dos dados do projeto SDSS.

Referências

Adelman-Mccarthy, J. et al. (2008). The Sixth Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, v. 175, n. 2, p. 297 – 313.

Ball, N. M., Brunner, R. J., Myers, A. D. (2006). Robust Machine Learning Applied to Astronomical Datasets I: Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. *The Astrophysical Journal*, v. 650, p. 497 – 509.

Bazell, D. e Aha, D. W. (2001). Ensembles of Classifiers for Morphological Galaxy Classification. *The Astrophysical Journal*, v. 548, p. 219 - 223.

Carvalho, R. R., Capelato, H.V e Campos Velho, H. F. (2008). Um Universo Escuro na Era da Tecnologia da Informação. *Boletim da Sociedade Brasileira de Astronomia* (submetido).

Hunt, E. B., Marin, J. e Stone, P. J. (1966). *Experiments in Induction*. New York: Academic Press.

Lin, N. e Thakar, A. R. (2008). CasJobs and MyDB: A Batch Query Workbench. *Computing in Science and Engineering*, v. 10, n. 1, p. 18 – 29.

Madsen, M., S. (1996). *The Dynamic Cosmos – Exploring the Physical Evolution of the Universe*. 1. ed. New York, NY, USA: Chapman & Hall, 144 p.

Murty, K. S., Kasif, S. e Salzberg, S. (1994). A System for Induction of Oblique Decision tree. *Journal of Artificial Intelligence Research*, v. 2, p. 1 – 32.

Petrosian, V. (1976). Surface brightness and evolution of galaxies, *The Astrophysical Journal*, v. 209, Nº 1.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, v.1, n. 1, p. 81 – 106.

Salzberg, S. et al. (1995). Decision Trees for Automated Identification of Cosmic Ray Hits in Hubble Space Telescope Images. *Publications of the Astronomical Society of the Pacific*, v. 107, p. 1 – 10.

Stoughton, C., Lupton, R.H., Bernardi, M. e Blanton, M. R. (2002). Sloan Digital Sky Survey: Early Data Release. *The Astrophysical Journal*, v. 123, p. 485 – 548.

Suchkov, A., Hanisch, R., J. e Margon, B. (2005). A Census of Object Types and Redshift Estimates in the SDSS Photometric Catalog from a Trained Decision Tree Classifier. *The Astronomical Journal*, v. 130, p. 2439 – 2452.

Szalay, A. S., Thakar, A. R. e Gray, J. (2008). The sqlLoader data-loading pipeline. *Computing in Science and Engineering*, v. 10, n. 1, p. 38 – 48.

Thakar, A. R., Szalay, A.S., Fekete, G. e Gray, J. (2008). The Catalog Archive Server Database Management System. *Computing in Science and Engineering*, v. 10, n. 1, p. 30 – 37.

Witten, I. H. e Frank, E. (2000). *Data mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*. San Francisco: Morgan Kaufmann, 371 p.

Zhang, Y. e Zhao, Y (2007). A Comparison of BBN, ADTree and MLP in Separating Quasars from Large Survey Catalogues. C*hinese Journal of Astronomy and Astrophysics*, v. 7, n. 2, p. 289 – 296.

Zhao, Y. e Zhang, Y. (2008). Comparison of Decision Tree Methods for Finding Active Objects. *Advances in Space Research*, v. 41, p. 1955 – 1959.