



20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Árvore de Decisão como Classificador Estrela/Galáxia

Eduardo Charles Vasconcellos¹, Reinaldo Ramos de Carvalho³,
Haroldo de Campos Velho², Hugo Vicente Capelato³, Reinaldo Roberto Rosa²

¹Programa de Mestrado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

²Laboratório Associado de Computação e Matemática Aplicada – LAC
Instituto Nacional de Pesquisas Espaciais – INPE

charles.edu@gmail.com

Abstract. *The Sloan Digital Sky Survey is the most important survey project nowadays. It release photometric data for approximately 130 thousands of celestial objects, and spectroscopic data for 1 thousand of them. This data must be classified and cataloged to be used in scientific research. The goal of this project is build a classifier to separate stars and galaxies in the Sloan data set, and generate a catalogue of this data.*

Resumo. *O Sloan Digital Sky Survey é o mais importante projeto de mapeamento celeste da atualidade. Ele disponibiliza dados fotométricos para aproximadamente 130 milhões de objetos celestes, e dados espectroscópicos para um milhão deles. Para que estes dados possam ser utilizados para estudar o universo, eles devem primeiro ser classificados e catalogados. Esse projeto tem por objetivo criar um classificador para separar estrelas de galáxias no conjunto de objetos do Sloan e gerar um catálogo próprio destes objetos.*

Palavras-chave: *methods: data analysis – catalogues – surveys – virtual observatory tools*

1. Introdução

Nas últimas décadas, a astronomia, assim como muitos outros campos da ciência, vem experimentando um enorme crescimento na quantidade, qualidade e complexidade de dados. Na astronomia, tal crescimento se deve aos novos detectores digitais que substituíram as antigas placas fotográficas levando a uma nova fase na observação do céu. Os mapeamentos digitais do céu tornaram-se comuns nas ultimas décadas, produzindo dados numa taxa de ~ 2 TB por noite, e em muitos casos esses dados precisam ser processados e catalogados quase que simultaneamente a sua obtenção.

O projeto de mapeamento celeste mais importante da atualidade é o Sloan Digital Sky Survey (SDSS). O SDSS já está no seu nono ano de operação com a cobertura de

mais de oito mil graus quadrados do céu e aproximadamente duzentos e trinta milhões de objetos detectados. Estão disponíveis para todos esses objetos dados de fotometria como magnitudes e elipsidade. O SDSS também fornece, para aproximadamente um milhão desses objetos, dados de espectroscopia. Esses dados observacionais (fotométricos e espectroscópicos) estão disponíveis no site do SDSS (<http://www.sdss.org/DR7/>) e são de livre acesso. Contudo, é preciso classificar esses objetos para que eles sejam cientificamente relevantes, ou seja, é preciso saber se o objeto observado é uma estrela ou uma galáxia.

Desde o final da década de 70, muitos trabalhos, como [3] e [4], tem buscado meios de classificar objetos observados em mapeamentos celestes. Em sua maioria esses trabalhos consistem em métodos paramétricos baseados em dados fotométricos e que só são eficientes para objetos muito brilhantes.

A partir de 1995 com o trabalho [9] técnicas baseadas no conceito de “machine learning” começaram a ser empregadas na busca por um melhor classificador estrela/galáxia para tratar os dados dos mapeamentos celestes. Mais recentemente, [1] classificou 143 milhões de objetos da terceira disponibilização de dados do SDSS utilizando como classificador uma árvore de decisão treinada com as magnitudes de 477.068 objetos do SDSS, cujos dados espectroscópicos estavam disponíveis.

Neste trabalho, temos por objetivo desenvolver uma ferramenta baseada em árvores de decisão ([6], [7] e [10]) para classificação estrela/galáxia de objetos em grandes bancos de dados. Neste primeiro momento estamos desenvolvendo uma árvore de decisão que será treinada com uma pequena fração dos objetos do SDSS (cerca de um milhão) que possuem uma classificação precisa com base em dados espectroscópicos. Ela será capaz de classificar os objetos observados pelo SDSS como estrelas ou galáxias com base em seus dados fotométricos. Essa classificação é fundamental para qualquer estudo científico tendo por base o SDSS. Na próxima fase do projeto estaremos aplicando o procedimento desenvolvido na fase anterior como um classificador estrela/galáxia que será incorporado ao projeto BRAVO (BRAZilian Virtual Observatory).

2. Conjunto de Dados Utilizado

O SDSS já está na sua sétima disponibilização de dados. Seu banco de dados possui dezenas de medidas observacionais (atributos) para cada objeto. Na fase atual do projeto, foram selecionados os atributos que apresentam maior relevância para classificação estrela/galáxia. Inicialmente foram escolhidos atributos que são representativos de uma certa classe, estrela ou galáxia, para objetos de alto brilho. Esses atributos são mostrados na tabela 1.

As magnitudes PSF, de Fibra, Petrosiana e Modelo são medidas do fluxo de energia emitido pelo objeto. A magnitude PSF consiste em um fluxo medido a partir de uma Point Spread Function (PSF). As estrelas, devido a suas dimensões e a sua distância a Terra, podem ser consideradas fontes pontuais, tornando essa magnitude particularmente adequada para medir fluxo estelar, já que uma PSF é um ajuste para dispersão luminosa sofrida por fontes pontuais.

A magnitude de fibra é a medida do fluxo do objeto através de uma fibra ótica com 3” de arco de diâmetro. Essa magnitude é relevante na separação de objetos com

Tabela 1. Atributos selecionados para classificação estrela/galáxia dos objetos SDSS.

Atributo	Atributo
Magnitude PSF	Raio contendo 90% do fluxo petrosiano
Magnitude de Fibra	Likelihood PSF
Manitude Petrosiana	Likelihood Exponencial
Magnitude Modelo	Likelihood DeVaucouleurs
Raio contendo 50% do fluxo petrosiano	Momentos Adaptativos

distâncias angularmente não resolvidas, ou seja, dois ou mais objetos que na imagem aparentam ser um único objeto.

A magnitude petrosiana é baseada em uma medida de fluxo proposta por [5]. Galáxias brilhantes aparecem como fontes extensas no céu. Essas galáxias tem formas e tamanhos distintos. Petrosian propôs uma medida de fluxo baseada na forma e tamanho do objeto, propiciando a medida de uma fração constante da luz do objeto (fluxo petrosiano), independente da posição e da distância do mesmo.

A magnitude modelo considera o fluxo do melhor ajuste entre dois perfis de brilho para uma galáxia. O SDSS calcula e disponibiliza o fluxo ajustado pelo perfil de DeVaucouleurs, equação (1), e pelo perfil exponencial, equação (2), sendo o melhor dentre os dois ajustes disponibilizado como magnitude modelo. Esses perfis de brilho são funções da distância ao centro do objeto, portanto caracterizam fontes extensas (galáxias).

$$I(r) = I_0 \exp 7.67 \left[\left(\frac{r}{r_e} \right)^{1/4} \right] \quad (1)$$

$$I(r) = I_0 \exp -1.68 \frac{r}{r_e} \quad (2)$$

Os raios petrosianos R50 e R90 são, respectivamente, os raios (distância ao centro do objeto) que contém 50% e 90% do fluxo petrosiano. Com esses valores calculamos o índice de concentração inverso dado por $c^{-1} = R50/R90$. Segundo [8] c^{-1} é um bom indicador de morfologia galática. Esse índice apresenta valores próximos 0.3 para objetos bem ajustadas pelo perfil de DeVaucouleurs devido a sua concentração de brilho na região central, e 0.43 para ajustes exponenciais devido à uniformidade de brilho. No caso de fontes pontuais esse valor fica próximo de 1 devido à fonte não ser angularmente resolvida.

O likelihood é a probabilidade de se atingir o valor medido do chi-quadrado para cada um dos modelos: DeVaucouleurs (deV_L), exponencial (exp_L) e PSF (star_L). Segundo a documentação do SDSS encontrada no site do projeto (<http://www.sdss.org/>), o likelihood fracional, que pode ser calculado para cada um dos modelos (equações 3, 4 e 5), pode ser um bom discriminante estrela/galáxia. A tabela 2 mostra um pequeno conjunto de objetos, com magnitudes entre 18 e 19, e seus valores de likelihood fracional.

$$f(\text{deV_L}) = \frac{\text{deV_L}}{\text{deV_L} + \text{exp_L} + \text{star_L}} \quad (3)$$

$$f(\text{exp_L}) = \frac{\text{exp_L}}{\text{deV_L} + \text{exp_L} + \text{star_L}} \quad (4)$$

$$f(\text{star_L}) = \frac{\text{star_L}}{\text{deV_L} + \text{exp_L} + \text{star_L}} \quad (5)$$

Tabela 2. Likelihood fracional.

Objeto	$f(\text{deV_L})$	$f(\text{exp_L})$	$f(\text{star_L})$
2	0.00	0.00	1.00
2	0.00	0.00	1.00
2	0.00	0.00	1.00
2	0.00	0.01	0.99
2	0.00	1.00	0.00
2	0.00	0.00	1.00
2	0.10	0.85	0.05
2	0.00	0.00	1.00
2	0.00	0.00	1.00
1	0.44	0.28	0.28
1	0.34	0.33	0.33
2	0.00	0.00	1.00

Os momentos aditivos são momentos de segunda ordem da intensidade do objeto. Eles são utilizados como uma medida de elipsidade. Espera-se que sejam bons discriminantes pelo fato de estrelas ser objetos com elipsidade muito baixa, enquanto galáxias apresentam valores de elipsidade mais elevados. Contudo, o SDSS provê outras medidas de elipsidade. No entanto, os momentos aditivos, dados pelas equações:

$$\langle col^2 \rangle = \frac{\sum [I(col, row) * w(col, row) * col^2]}{\sum [I * w]}, \quad (6)$$

$$m_{rr_cc} = \langle col^2 \rangle + \langle row^2 \rangle, \quad (7)$$

$$m_{e1} = \frac{\langle col^2 \rangle - \langle row^2 \rangle}{m_{rr_cc}}, \quad (8)$$

$$m_{e2} = \frac{\langle col^2 \rangle * \langle row^2 \rangle}{m_{rr_cc}}, \quad (9)$$

onde col e row são as coordenadas de um pixel no detector, $I(col, row)$ é a intensidade no pixel e $w(col, row)$ é uma função peso [2], são as medidas mais confiáveis para objetos com valores mais elevados de magnitude.

Os atributos selecionados serão adquiridos do banco de dados do SDSS. No total, nosso conjunto de dados será composto por esses atributos para aproximadamente 130 milhões de objetos com valores de magnitude até 25.

3. Árvore de Decisão

Uma árvore de decisão é uma estrutura simples formada por nós e ramos, como mostrada na figura 1, que é gerada a partir de um conjunto de dados. Cada nó da árvore representa um atributo do conjunto de dados, e cada ramo os possíveis valores para aquele atributo. A estrutura da árvore é top-down (de cima para baixo), sendo o nó do topo chamado de raiz e cada um dos nós dos quais não saem ramos de folha. A árvore nada mais é do que um conjunto de regras para separação em classes dos dados presentes no conjunto. É chamado de regra o caminho que parte da raiz e, passando por ramos e nós, chega a uma folha. Uma folha é a representação de uma determinada classe dentro do conjunto de dados. Uma única classe pode ser representada por mais de uma folha, ou seja, pode haver mais de uma regra classificatória para aquela classe.

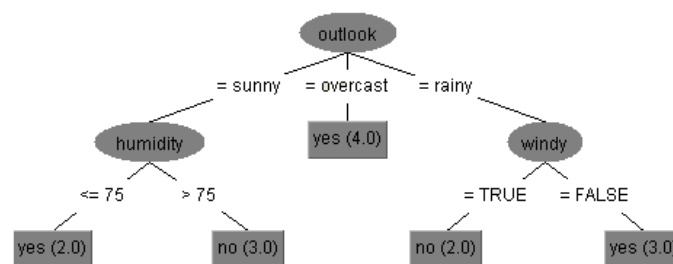


Figura 1. Uma árvore de decisão simples baseada no problema modelo weather ([10]). A árvore foi gerada com a ferramenta Weka.

Existem diversos algoritmos para geração de árvores de decisão ([10]), sendo que o ID3 desenvolvido por [6] e atualizado para o C4.5 em 1993 ([7]) é um dos mais populares.

A utilização de árvores de decisão não se restringe apenas a descobrir classes dentro de um determinado conjunto de dados. As regras classificatórias devem poder ser generalizadas de maneira a classificar dados com mesmo formato mas fora do conjunto usado na geração da árvore. Esse objetivo pode ser alcançado com um conjunto de treinamento suficientemente representativo ou com a simplificação da árvore ([10]) mantendo apenas as regras mais genéricas. O algoritmo C4.5 conta com um sistema de poda para executar o processo de simplificação da árvore por ele gerada.

No momento atual do projeto, um conjunto de dados contendo os atributos descritos na seção anterior para 1,049,304 objetos do SDSS, cuja classificação estrela/galáxia com base em espectroscopia é precisa, será utilizado para gerar uma árvore de decisão através da implementação Weka do algoritmo C4.5. Nesta primeira etapa estamos interessados somente na separação em classes dos objetos do próprio conjunto. Futuramente, essa árvore será aplicada aos demais objetos do SDSS. A necessidade da generalização da árvore através de um processo de poda ou a manutenção da árvore completa só será verificada através da execução de uma bateria de testes, que serão executados com auxílio da ferramenta Weka.

A ferramenta Weka está sendo utilizada com intuito de otimizar o uso do reduzido período de tempo disponível para execução do projeto. A Weka foi desenvolvida com base na obra [10] pelos próprios autores, e está implementada na linguagem Java, disponibilizando uma interface gráfica simples para o usuário. Ela possui, além dos de árvore de decisão, diferentes algoritmos baseados no conceito de “machine learning” para mineração

de dados. A ferramenta possui código aberto e permite a possibilidade de sua utilização como biblioteca Java.

4. Resultados Preliminares e Perspectivas Futuras

Um conjunto preliminar de dados contendo $\sim 5\%$ (50,000 objetos) do conjunto de dados inicial (1,049,304 objetos) foi utilizado para gerar uma árvore teste. Como esses dados preliminares são um subconjunto dos dados iniciais, ele também possui classificação precisa. Aplicando a implementação C4.5 do Weka para gerar uma árvore não podada utilizando a árvore resultante para classificar o próprio conjunto gerador, foi obtidos 97% de completude (galáxias classificadas corretamente) e 0.2% de contaminação (estrelas classificadas como galáxias).

Os objetos do conjunto preliminar possuem magnitudes na banda¹ r entre 18 e 19. Esse intervalo de magnitudes contém o início da região, dentro do espaço de parâmetros, onde estrelas e galáxias ocupam posições muito próximas, como podemos ver na figura 2. Portanto, o resultado obtido no estudo preliminar indica que os atributos escolhidos podem gerar um bom discriminador estrela/galáxia para os objetos do SDSS.

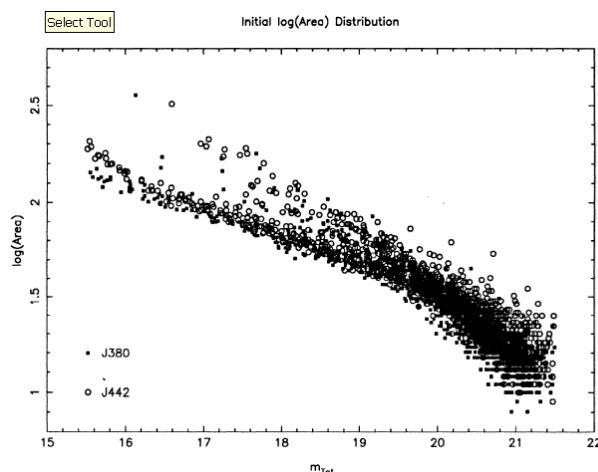


Figura 2. Gráfico bidimensional do espaço de parâmetros. J380 e J442 são dois tipos de placa fotográfica. Em magnitudes inferiores a 18 podemos ver claramente a distinção entre estrelas e galáxias. As estrelas ocupam a região inferior. A partir de magnitude 18, estrelas e galáxias começam a ocupar a mesma região do espaço (Fonte:[9]).

Os 13 algoritmos do WEKA foram testados e os resultados são mostrados na Tabela 3. O FT mostrou-se como sendo o algoritmo de melhor desempenho e foi utilizado para gerar uma árvore utilizada na classificação e catálogo de todos os objetos SDSS com magnitudes $14 \leq r \leq 21$. Nosso catálogo contém ~ 69 milhões de objetos e apresenta um percentual de galáxias classificadas corretamente entrono de 99.5% para magnitudes $14 \leq r < 19$, e 85% para os demais objetos.

¹Sistema de filtro ugriz utilizado nas observações do SDSS. A banda é a função resposta do detector em um certo comprimento de onda.

Tabela 3. Resultados para o estudo dos algoritmos do WEKA. Primeira coluna mostra o nome do algoritmo; a segunda, a porcentagem de galáxias classificadas corretamente no intervalo de magnitudes [14, 19]; a terceira, porcentagem de galáxias classificadas corretamente no intervalo de magnitudes [19, 21]

Algorithm	$\langle \text{Compl} \rangle_{\text{bright}}$	$\langle \text{Compl} \rangle_{\text{faint}}$
Decision Stump	99.20(± 0.17)	68.06(± 1.20)
NBTree	99.64(± 0.16)	79.19(± 1.39)
J48graft	99.74(± 0.12)	80.93(± 1.16)
Simple Cart	99.63(± 0.16)	81.56(± 1.13)
J48	99.73(± 0.12)	81.70(± 0.96)
REPTree	99.50(± 0.18)	82.76(± 1.09)
Random Tree	99.50(± 0.18)	82.76(± 1.09)
Random Forest	99.77(± 0.12)	83.15(± 1.14)
BFTree	99.69(± 0.15)	83.18(± 1.10)
ADTree	99.73(± 0.12)	83.80(± 1.12)
LMT	99.66(± 0.15)	83.91(± 1.14)
LADTree	99.70(± 0.14)	84.39(± 1.10)
FT	99.64(± 0.15)	84.98(± 1.08)

Referências

- [1] Ball, N. M., Brunner, R. J., Myers, A. D., and Tchong, 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*, 650:497-509.
- [2] Bernstein, G. M. and Jarvis, M. (2002). Shapes and shears, stars and smears: Optimal measurements for weak lensing. *The Astronomical Journal*, 123:583618.
- [3] Heydon-Dumbleton, N. H., Collins, C. A., and MacGillivray, H. T. (1989). The edinburgh/durham southern galaxy catalogue. ii - image classification and galaxy number counts. *Royal Astronomical Society, Monthly Notices*, 238:379406.
- [4] Maddox, S. J., Efstathiou, G., Sutherland, W. J., and Loveday, J. (1990). The apm galaxy survey. i - apm measurements and star-galaxy separation. *Royal Astronomical Society, Monthly Notices*, 243:692712.
- [5] Petrosian, V. (1976). Surface brightness and evolution of galaxies. *Astrophysical Journal*, 209:L1-L5.
- [6] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81-106.
- [7] Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers.
- [8] Shimasaku, K., Fukugita, M., Doi, M., Hamabe, M., Ichikawa, T., Okamura, S., Sekiguchi, M., Yasuda, N., Brinkmann, J., Csabai, I., Ichikawa, S.-I., Ivezić, Z., Kunszt, P. Z., Schneider, D. P., Szokoly, G. P., Watanabe, M., and York, D. G. (2001). Statistical properties of bright galaxies in the sloan digital sky survey photometric system. *The Astronomical Journal*, 122:1238-1250.

- [9] Weir, N., Fayyad, U. M., and Djorgovski, S. (1995). Automated star/galaxy classification for digitized poss-ii. *Astronomical Journal*, 109:2401.
- [10] Witten, I. H. and Frank, E. (2000). *Data Mining: pratical machine learning toos and techniques with java implementations*. Morgan Kaufmann Publishers.