Árvore de Decisão Usada como Discriminador entre Estrela e Galáxia

Eduardo Charles Vasconcellos
Computação Aplicada, LAC, INPE,
12227-010, São José dos Campos, SP
E-mail: charles.edu@gmail.com,

Reinaldo Ramos de Carvalho **Hugo Vicente Capelato**
Divisão de Astrofísica- INPE
12227-010, São José dos Campos, SP
E-mail: rrdecarvalho2008@gmail.com, hugo@das.inpe.br,

Haroldo Fraga de Campos Velho **Reinaldo Roberto Rosa**
Laboratório Associado de Computação e Matemática Aplicada- INPE
12227-010, São José dos Campos, SP
E-mail: haroldo@lac.inpe.br, reinaldo@lac.inpe.br.

RESUMO

Abril de 2009

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 os de Heydon-Dumbleton et al. (1989) [2] e Maddox et al. (1990) [3], 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 de Weir et al. (1995) [7] 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, Ball et al. (2006) [1] classificou 143 milhões de objetos da terceira disponibilização de dados do SDSS uti-

lizando 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 criar uma ferramenta para analisar e catalogar objetos (estrelas e galáxias) detectados nas imagens da sétima disponibilização de dados do SDSS. Essa ferramenta que estamos desenvolvendo consiste em uma árvore de decisão ([4], [5] e [6]) 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. Por exemplo, para o estudo de estruturas em grande escala no universo, utilizando aglomerados de galáxias, é preciso saber quais dos objetos observados na região celeste de interesse são galáxias.

O banco de dados do SDSS possui dezenas de medidas observacionais (atributos) para cada objeto. Na fase atual do projeto, estamos selecionando os atributos que sejam relevantes a classificação estrela/galáxia. Inicialmente foram escolhidos atributos que são representativos de uma certa classe, estrela ou galáxia, para objetos de alto brilho. Testes preliminares com um conjunto de 50 mil objetos de classificação bem conhecida, entre magnitudes 18.0 e 19.0, foram realizados com auxílio da ferramenta Weka [6]. Nesses testes foi utilizado o algoritmo J48 [6] para construção da árvore de decisão e o algoritmo de poda C4.5 [5] com fator de confiança 0.25 para otimizar o classificador. Esses testes resultaram em uma completeza (galáxias classificadas corretamente) de aproximadamente 98% e uma contaminação (estrelas classificadas como galáxias) de cerca de 10%.

Esperamos ter um classificador operacional em agosto e um catálogo próprio dos objetos observados pelo SDSS no final do ano. Futuramente este classificador estrela/galáxia será implementado em um conjunto de ferramentas para processamento, análise e catalogação de imagens astronômicas que serão disponibilizadas pelo projeto BRAVO (BRAzilian Virtual Observatory). O BRAVO estará disponível para a comunidade astronômica brasileira como uma ferramenta moderna para lidar com a nova fase da produção de dados.

Palavras-chave: *SDSS, Árvore de Decisão, Classificação estrela/galáxia*

Referências

[1] N. M. Ball; R. J. Brunner; A. D. Myers, 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 (2006) 497-509.

[2] N. H. Heydon-Dumbleton; C. A. Collins; H. T. Macgillivray, The edinburgh/durham southern galaxy catalogue. ii - image classification and galaxy number counts, *Royal Astronomical Society, Monthly Notices*, 238 (1989) 379-406.

[3] S. J. Maddox; G. Efstathiou; W. J. Sutherland; J. Loveday, The apm galaxy survey. i - apm measurements and star-galaxy separation. *Royal Astronomical Society, Monthly Notices*, 243 (1990) 692-712.

[4] J. Quinlan, Induction of Decision Trees, *Machine Learning*, 1 (1986) 81-106

[5] J. Quinlan, "C4.5: Programs for machine learning", Morgan Kaufmann Publishers, 1993.

[6] I. H. Witten; E. Frank, "Data Mining: pratical machine learning toos and techniques with java implementations", Morgan Kaufmann Publishers, 2000.

[7] N. Weir; U. M. Fayyad; S. Djorgovski, Automated Star/Galaxy Classification for Digitized Poss-II, *Astronomical Journal*, 109 (1995) 2401-2414.