Árvores de decisão em classificação de dados astronômicos

Renata S. R. Ruiz, Haroldo F. de Campos Velho, Rafael D. C. dos Santos
Laboratório de Computação e Matemática Aplicada, LAC, INPE
12245-970-000, São José dos Campos, SP
E-mail: [renata, haroldo, rafael.santos]@lac.inpe.br

Marina Trevisan
Depto. de Astronomia, IAG, USP
05508-090, São Paulo, SP
E-mail: trevisan@astro.iag.usp.br

Resumo: *Os registros de astronomia ótica constituem uma fonte de informação extremamente importante. Um desafio científico relevantes é separar estrelas de galáxias a partir de dados fotométricos. Realizar esta classificação é o objetivo deste trabalho. árvores de decisão é a técnica empregadaneste artigo. Dados do projeto \emph{Sloan Digital Sky Survey} foram usados para treinar e testar a metodologia. A amostra final é composta por 43289 estrelas e 452400 galáxias, onde foram selecionados os par\^ametros fotométricos considerados relevantes na distinção entre estrelas e galáxias. O método apresentou bons resultados, mostrando-se uma técnica competitiva.*

1. Introdução

O entendimento sobre o Cosmos tem sido alterado ao longo dos tempos. Na cultura ocidental, antes da era moderna, o ponto de vista aristotélico prevalecia: existia uma física para as coisas da Terra e outra física para os corpos celestiais. O inglês Issac Newton altera para sempre este paradigma e desenvolve 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.

O século XX vai desafiar o modelo Newtoniano. Uma nova interpretação vai surgir com o desenvolvimento da teoria da relatividade de Albert Eistein. Em 1917, para acomodar a teoria da relatividade geral (trabalho de 1915) com a idéia de um Universo estático, Einstein introduziu um novo termo em suas equações: a Constante Cosmológica (Λ). Em 1929, o astrônomo americano Edwin Powel Hubble publica um resultado de que as galáxias estão se afastando, independente de que direção esteja sendo realizada a observação. Ou seja, o Universo está em expansão. Isso inicia uma nova era na ciência.

No final da década de 90, os cosmólogos descobriram que o universo está numa expansão *acelerada* (e não desacelerada, como se acreditava até muito recentemente). Para explicar esta observação sugeriu-se a introdução de um conceito: *energia escura*. Esta seria uma forma de energia **gravitacionalmente repulsiva** (os cálculos indicam que esta é forma mais abundante de energia no Universo). A pesquisa em energia escura é um dos temas de intensa investigação na cosmologia e astronomia de hoje.

A gravitação faz surgir uma hierarquia de estruturas cosmológicas. Aglomerados de galáxias são estruturas formadas por dezenas, centenas, ou milhares de galáxias ligadas pela gravitação. A din\^amica composta pela força da gravidade e pela expansão do Universo descreve a história da formação destas estruturas (aglomerados, super-aglomerados e assim por diante).

O estudo das grandes estruturas existentes no universo depende do correto mapeamento de galáxias. As grandes bases de dados astronômicos existentes hoje fornecem uma possibilidade de estudo de grandes estruturas sem precedentes.

Porém, há um problema prático da astronomia observacional que tem uma importante ligação com a pesquisa em energia escura: numa imagem astronômica, nem sempre é fácil distinguir entre uma galáxia e

uma estrela. A dificuldade é devida porque quanto mais distante está uma galáxia do nosso planeta, menor é o registro (tamanho na imagem) desta galáxia e menor é a luminosidade observada. Quando se atingem 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. A figura 1. (ver referência [2]) ilustra a dificuldade mencionada.

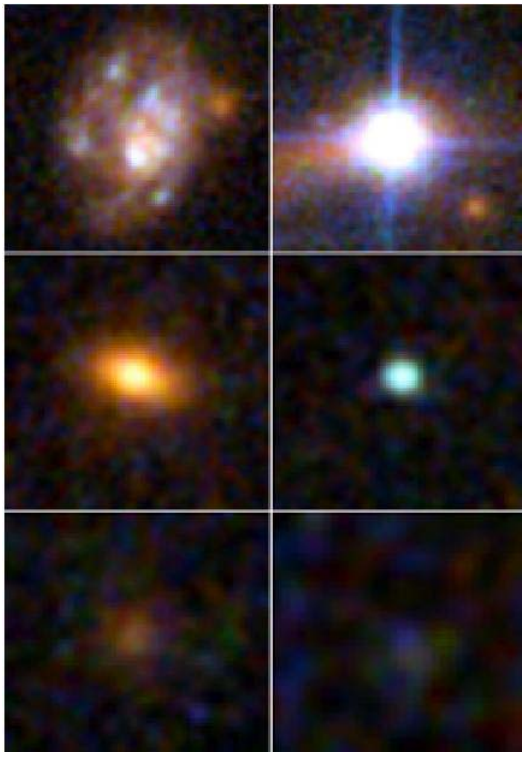

Figura 1: Separação galáxia-estrela: no painel superior é simples realizar a identificação de um objeto de alta luminosidade; no painel central a identificação ainda é simples, mas é muito difícil de ser realizada de forma não automática; no painel inferior a identificação é muito complexa.

Quando conhecidas as distâncias dos objetos observados, se torna fácil a distinção entre estrelas e galáxias, uma vez que galáxias estão sempre a distâncias bem maiores que estrelas. Essas distâncias podem ser medidas com grande precisão a partir dos espectros dos objetos. Além disso, espectros de galáxias e de estrelas são bem distintos entre si, e isso também pode ser usado na classificação. Porém, a aquisição de espectros em geral requer um maior tempo de observação e utilizar dados fotométricos torna as observações mais eficientes. 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.

2. Árvores de decisão para problemas de classificação

Considere um conjunto de objetos que são descritos em termos de uma coleção de atributos. Esses objetos podem pertencer a diferentes classes. Cada atributo mede alguma característica importante de um objeto.

Agora considere um conjunto de treinamento, cuja classe de cada objeto é conhecida. Se o conjunto de treinamento contém dois objetos que têm valores idênticos para dado atributo e mesmo assim pertencem a classes diferentes, é impossível diferenciar esses objetos somente considerando tal atributo. Neste caso, considera-se que este atributo é inadequado para o conjunto de treinamento e também para a tarefa de indução. A tarefa de indução é 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 outros nodos representam atributos baseados em testes com um ramo para cada possível saída. 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 até uma folha ser encontrada. Em tal caso garante-se que o objeto pertence a classe nomeada pela folha.

Se os atributos são adequados é sempre possível construir uma árvore de decisão que classifica corretamente cada objeto no conjunto de treinamento e normalmente existem muitas árvores de decisão corretas. A essência da indução é ir além do conjunto de treinamento, isto é, classificar corretamente outros objetos. Para conseguir isto a árvore de decisão deve capturar alguma relação significativa entre a classe do objeto e os valores de seus atributos. Quando tem-se duas árvores de decisão que classifica corretamente um conjunto de treinamento, deve-se escolher

a mais simples, uma vez que, ela é mais adequada para capturar a estrutura inerente do problema e assim, vai classificar corretamente mais objetos fora do conjunto de treinamento.

São diversos os algoritmos de indução de árvore de decisão conhecidos na literatura. O algoritmo ID3, desenvolvido por [5], é um dos algoritmos mais populares na área da indução de árvore de decisão.

A idéia básica do ID3 é iterativa. Um subconjunto do conjunto de treinamento chamado janela é escolhido aleatoriamente e uma árvore de decisão é formada a partir dele. Todos os outros objetos do conjunto de treinamento são classificados usando a árvore. Se esta árvore fornecer a resposta correta para todos os objetos o processo termina, se não, uma seleção dos objetos classificados incorretamente é adicionada a janela e o processo continua.

O cerne do problema é como formar uma árvore de decisão para uma coleção arbitrária de C objetos. Se C é vazio ou contém somente objetos de uma classe, a árvore de decisão mais simples é justamente uma folha classificada com aquela classe. Caso contrário, seja T qualquer 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 $\{C_1, C_2, \ldots, C_w\}$, de C, com C_i contendo aqueles objetos que tem saída O_i. No pior caso essa estratégia fornecerá subconjuntos de um único objeto, que satisfaz a exigência de uma classe por folha. Assim, uma vez que um teste pode sempre ser encontrado de uma divisão não trivial de qualquer conjunto de objetos, este procedimento sempre permite obter uma árvore de decisão que classifique corretamente os objetos em C [5].

A escolha do teste é crucial para a árvore de decisão ser simples. O ID3 adota uma informação baseada no método que depende de duas hipóteses:

H1: Toda árvore de decisão correta para C classificará objetos na mesma proporção que sua representação em C. 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 $n/(p+n)$.

H2: Quando uma árvore de decisão é usada para classificar um objeto, ela retorna uma classe. Árvore de decisão pode assim ser considerada como uma fonte de mensagem P ou N com a informação necessária prevista para gerar a mensagem dada por:

$$I(p,n) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) \\ \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \qquad (1)$$

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

$$E(A) = \sum_{i=1}^{v}\left[\frac{p_i + n_i}{p+n}\right] I(p_i, n_i) \qquad (2)$$

em que o peso para o i-ésimo ramo é proporcional aos objetos em C que pertencem C_i.

A informação ganha pela ramificação sobre A é, desta forma:

$$G(A) = I(p,n) - E(A). \qquad (3)$$

O algoritmo ID3 examina todos os atributos candidatos e escolhe A para maximizar o ganho de A, forma as árvores como acima e então usa o mesmo processo recursivamente para formar a árvore de decisão para os subconjuntos restantes, C_1, C_2, ... , C_v. As primeiras versões eram limitadas a parâmetros discretos, mas o agoritmo sofreu uma significativa evolução com a introdução de parâmetros contínuos na versão [6]. O pacote *Weka classifier* tem uma versão do C4.5 conhecida como J.48.

3. Resultados de classificação

Os dados utilizados neste trabalho são os dados de seis anos do projeto *Sloan Digital*

Sky Survey (SDSS) (http://www.sdss.org/) [1,2]. Este levantamento cobre uma área de ≈10000 graus quadrados do céu, contendo imagens de 287 milhões de objetos em cinco bandas fotométricas *u*, *g*, *r*, *i*, *z*. Além disso, há também o levantamento espectroscópico, que cobre uma área de ≈7500 graus quadrados, com mais de um milhão de objetos catalogados. Como mencionado acima, 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 a serem utilizadas neste trabalho.

Cada objeto identificado nas imagens é classificado pelo próprio *pipeline* do SDSS, em todas as cinco bandas. O mesmo é feito para os dados espectrosóspicos, de forma que há cinco parâmetros. Os parâmetros relacionados com esta classificação são *type-u*, *type-g*, *type-r*, *type-i*, *type-z* e *SpecClass*. Os cinco primeiros recebem os valores 3 se for galáxia e 6 se for estrela e *SpecClass*=1 no caso de estrela e *SpecClass*=2 se galáxia. Outras classificações são possíveis, mas isso não é relevante neste trabalho. Baseando-se nisso, o primeiro critério de seleção foi exigir que o objeto tenha a mesma classificação nestes seis parâmetros simultaneamente. Isso reduz a quase nula a probabilidade de ter na amostra um objeto classificado erroneamente (1,5% dos objetos classificados nas bandas *u*, *g*, *r*, *i* e *z* como galáxias não possuem a mesma classificação quando considerado o espectro).

Entre estes objetos seguramente classificados como estrelas e galáxias, foram ainda impostas restrições aos *flags* de qualidade, gerados pelo *pipeline* do SDSS, relacionados à saturação do objeto e à detecção de múltiplos picos de intensidade nas imagens. A amostra final é composta por 43.289 estrelas e 45.2400 galáxias. Os dados foram obtidos através do servidor CasJobs do SDSS [3, 9, 10, 11].

Dos objetos da amostra final foram selecionados os parâmetros fotométricos que foram considerados relevantes na distinção entre estrelas e galáxias, em todas as cinco bandas.

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 estão na tabela 7 de Stoughton [7]. *nprof* corresponde ao números de anéis para os quais ainda existe um sinal mensurável.

PetroR50, PetroR90: Para cada objeto é definido o perfil de brilho superficial Petrosiano [4], e a partir deste são definidos os raios PetroR50 e PetroR90, 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. Assim, uma informação interessante obtida através desses atributos é a excentricidade, já que estrelas têm forma circular.

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 [7].

***Redshift* espectroscópico:** Este não é um dado fotométrico, e sim obtido a partir dos espectros. Como é baseado em linhas de emissão e absorção, e não apenas no fluxo em bandas, é uma medida mais precisa de distância do que o *redshift* fotométrico. O *redshift* de um objeto é medido como o deslocamento relativo do comprimento de onda emitido pela fonte e observado:

$$z = \frac{\lambda_{Observado} - \lambda_{emitido}}{\lambda_{emitido}} \qquad (4)$$

o aumento no comprimento de onda é causado pela expansão do universo: quanto mais distante o objeto, maior é o seu *redshift* (*z*). Apesar de esperar-se que estrelas tenham sempre o *redshift* nulo, isso nem sempre é

verdade, pois a mudança no comprimento de onda também pode ser causada por efeito Doppler devido ao movimento da fonte em relação ao observador.

Da amostra de dados do SDSS, composta por 43.289 estrelas e 45.2400 galáxias, o treinamento (criação da árvore) foi realizado com 631 estrelas e 4369 galáxias (total 5000 objetos). Os dados utilizados para testar o desempenho das árvores geradas contêm 925 estrelas e 9075 galáxias (total 10000 objetos).

Após definir os atributos e selecionar os dados para treinamento e teste para a classificação, foram criadas as árvores de decisão utilizando o software *Weka* por meio do algoritmo J4.8. Para configurar a primeira árvore, foram utilizados todos os atributos. O algoritmo J.48, selecionou de forma automática apenas o atributo *z* (*redshift*) para realizar a classificação. Esta estrutura de árvore é bem simples e não exigiu um tempo de processamento elevado. A figura 4.1 mostra essa árvore de decisão. A árvore criada estabeleceu um valor crítico para *z*, para o qual $z \leq 0{,}001546$ o objeto é classificado como estrela (se $z > 0{,}001546$ o objeto é uma galáxia).

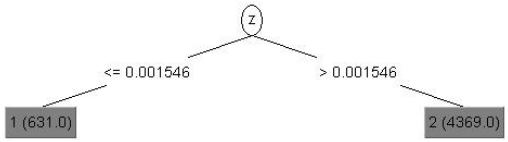

Figura 4.1: Árvore gerada com base apenas no atributo *redshit* (*z*).

Esse arquivo de teste conta com 10.000 objetos e sua classificação é conhecida. Comparando a classificação obtida pelo modelo e a classificação real, o modelo teve um acerto de 100%, confirmando o bom desempenho da árvore gerada. A matriz de confusão (*confusion matrix*) para árvore dada pela figura 4.1 é exibida na tabela 4.1. A matriz de confusão indica quais instâncias foram classificadas de forma correta e incorreta. Se todo elemento fora das diagonais é igual a zero então se tem uma classificação 100% correta.

Tabela 4.1: Matriz de confusão para 1ª árvore gerada.

a	b	Classificação
925	0	a=estrela
0	9075	b=galáxia

Eventualmente o parâmetro *z* pode não ser conhecido, ou conter imprecisões. Desse modo, criou-se uma outra árvore de decisão com o parâmetro z removido do conjunto de atributos. No caso desta árvore, construída a partir de 20 parâmetros (4 atributos x 5 bandas), foi configurada para, no mínimo, classificar 10 objetos por folha. A figura 4.2 exibe a nova árvore de decisão criada. Com o mesmo procedimento adotado na árvore anterior, realizou-se a aplicação desta árvore ao conjunto de teste. Como pode ser observado na matriz de confusão dada na tabela 4.2, a nova árvore também mostrou-se bastante eficiente para realizar a classificação, obtendo um índice de 99.48% de acertos.

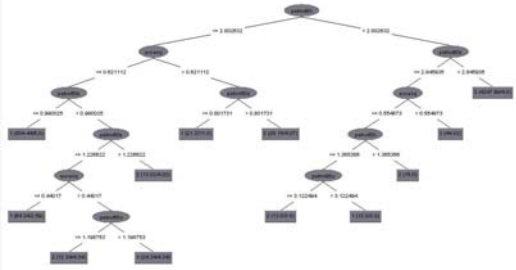

Figura 4.2: Árvore gerada sem o parâmetro *z*.

Tabela 4.2: Matriz de confusão para 2ª árvore gerada.

a	b	Classificação
923	2	a=estrela
50	9025	b=galáxia

4. Conclusões

Foram criadas duas árvores de decisão utilizando a ferramenta *Weka*, por meio do algoritmo J.48, com o objetivo de efetuar a classificação de objetos astronômicos em estrelas e galáxias. Duas árvores foram estruturadas para realizar a classificação. As tabelas 4.1 e 4.2 mostram que os resultados foram bastante satisfatórios, como se pode observar através das matrizes de confusão.

A primeira árvore, embora bastante simples, obteve um índice de acerto de 100% sobre os dados de teste. Vale lembrar que os dados utilizados para o teste são diferentes dos dados

utilizados para o treinamento. Na segunda árvore foi realizado um processo de poda (limite mínimo de classificação por folha) para eliminar os nós e as folhas não significativas. Essa árvore também apresentou um resultado muito bom com um erro de apenas 0,52%. Mediante os resultados obtidos podemos concluir que o método de árvore de decisão, baseado no algoritmo J.48, mostrou-se eficiente para classificação de dados astronômicos, como os atributos indicados. A ferramenta *Weka* mostrou eficaz para uma primeira avaliação ou exploração da técnica.

Referências

[1] Adelman-McCarthy, J., Agueros, M.A., Allam, S.S., et al., The Sixth Data Release of the Sloan Digital Sky Survey. The Astrophysical Journal Supplement Series, 175, 297, (2008).

[2] R. R. Carvalho, H. V. Capelato, H. F. de Campos Velho, Um Universo Escuro na Era da Tecnologia da Informação, Boletim da Sociedade Brasileira de Astronomia www.sab-astro.org.br/boletim.html, (2008) – submetido.

[3] N. Lin, A. R. Thakar, A.R., Computing in Science and Engineering, **10**(1), 18—29, (2008).

[4] V. Petrosian, Surface brightness and evolution of galaxies. The Astrophysical Journal, **209**, L1, (1976).

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

[6] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufman, 1993.

[7] C. Stoughton, R.H. Lupton, M. Bernardi, M.R. Blanton, M. R., et al., The Astrophysical Journal, **123**, 485. (2002).

[8] M.A. Strauss, D.H. Weinberg, R.H. Lupton et al. The Astronomical Journal, **124**(3), 1810—1824, (2002).

[9] A.S. Szalay, A.R. Thakar, J. Gray, Computing in Science and Engineering}, **10**(1), 38-48, (2008).

[10] A.R. Thakar, A.R. \textit{Computing in Science and Engineering}, **10**(1), 9--12. (2008).

[11] A.R. Thakar, A.S. Szalay, G. Fekete, J. Gray, Computing in Science and Engineering, **10**(1), 30-37, (2008).

[12] D.G. York, J. Adelman, J.E. Anderson, J.E., et al. (mais 144 nomes), The Astronomical Journal, **120**(3), 1579-1586, (2000).