TEMA Tend. Mat. Apl. Comput., **10**, No. 1 (2009), 75-86.

Árvores de Decisão na Classificação de Dados Astronômicos

R.S.R. RUIZ[1], H.F. DE CAMPOS VELHO [2], R.D.C. SANTOS [3], Laboratório Associado de Computação e Matemática Aplicada, LAC, INPE, 12227-010 São José dos Campos, SP, Brasil.

M. TREVISAN[4], Departamento de Astronomia, IAG, USP, 05508-900, São Paulo, SP, Brasil.

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) e sua aplicação na construção de classificadores baseados em atributos fotométricos para classificar objetos astrônomicos em estrelas e galáxias. Dados do projeto Sloan Digital Sky Survey (SDSS) foram utilizados 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.

Palavras-chave. Árvore de decisão, dados astronômicos, parâmetros fotométricos.

1. Introdução

O entendimento sobre a origem e evolução do Universo tem sido alterado ao longo dos tempos. Anteriormente, prevalecia a visão aristotélica: existia uma física para os fenômenos da Terra e outra física para os corpos celestiais. Isaac Newton alterou para sempre este paradigma e desenvolveu um modelo físico-matemático constituído de poucos postulados e algumas leis, como a formulação matemática da gravitação Universal. O modelo cosmológico de Newton é de um Universo infinito aparentemente estável e estático.

Em 1917, Albert Einstein propôs um modelo cosmológico relativístico, considerando ainda 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 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 [7, 14].

[1]renata@lac.inpe.br - A autora agradece a FAPESP bolsa de doutorado (2007/54133-0)
[2]haroldo@lac.inpe.br
[3]rafael.santos@lac.inpe.br
[4]marinatrevisan@gmail.com

Conforme [14], 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 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. As dificuldades para análise baseada em atributos fotométricos estão relacionadas a vários fatores, entre eles: baixa luminosidade, baixo brilho superficial, perfis extensos, diferentes resoluções angulares e a razão sinal-ruído. Por exemplo, 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 [7].

A Figura 1 ilustra o tipo de dificuldade referente a luminosidade, na figura pode-se observar que na parte superior é simples realizar a classificação de um objeto, na parte central a classificação 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.

Neste contexto, separar estrelas de galáxias a partir de dados fotométricos é um desafio interessante e o objetivo deste trabalho é aplicar a técnica de árvores de decisão a este problema. Há na literatura uma série de trabalhos utilizando árvores de decisão para classificar objetos astronômicos [4, 20, 26, 27]. Outras técnicas de classificação também muito utilizadas são as redes neurais artificiais [2, 8, 15] e o algoritmo *Friends of Friends* [10].

Em particular, para dados do projeto *Sloan Digital Sky Survey* (SDSS) uma abordagem em árvores de decisão foi utilizada por [22] 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 [16].

Árvore de decisão também foi utilizada por [3] 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 em cerca de 143 milhões de objetos do projeto, sendo este o primeiro trabalho a utilizar árvores de decisão para classificar um conjunto inteiro de dados do SDSS.

No presente estudo, será utilizado algoritmo de construção de árvore de decisão C4.5 [18], implementado no software *Waikato Environment for Knowledge Analysis* (WEKA) como J4.8 [25], para desenvolver classificadores baseados em atributos

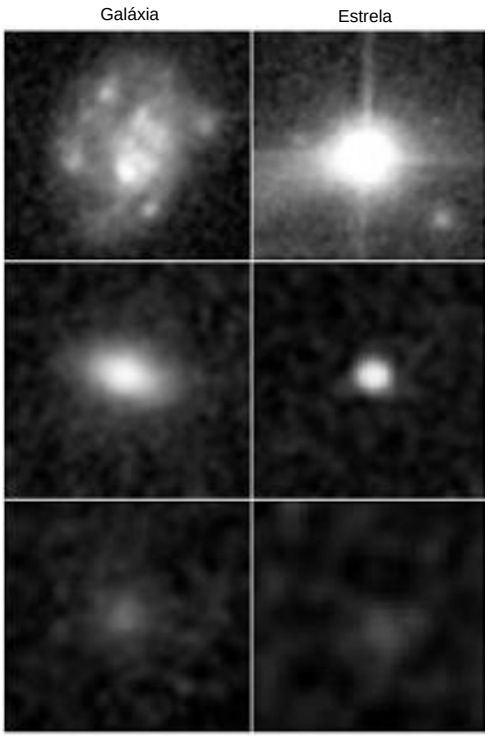

Figura 1: Exemplos de dificuldade crescente de separação estrela-galáxia – Fonte: [7].

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. Árvores de Decisão e Classificação

Dado um conjunto de objetos descritos em termos de uma coleção de atributos, estes objetos podem pertencer a diferentes classes. Cada atributo expressa alguma característica importante de um objeto. Parte destes objetos, serão considerados para o treinamento e tem sua classificação previamente conhecida. Conforme [19] 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 estru-

tura simples em que as folhas contêm as classes, os nodos não-folhas representam atributos baseados em testes com um ramo para cada possível saída [11, 18, 19]. Para classificar um objeto, começa-se com a raiz da árvore, aplica-se o teste em cada nodo 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.

Com os atributos adequados, é sempre possível construir uma árvore de decisão que classifique corretamente os objetos no conjunto de treinamento e normalmente existem muitas árvores de decisão corretas. Mas o objetivo dos algoritmos de indução (construção) é ir além do conjunto de treinamento, isto é, criar árvores capazes de classificar corretamente outros objetos. Para conseguir isto, tais algoritmos devem capturar alguma relação significativa entre a classe do objeto e os valores de seus atributos.

São diversos os algoritmos de indução de árvores de decisão conhecidos na literatura, dos quais destacam-se: *Random Forest* [5, 6],*ADTree* [9], *NBTree* [12], C4.5 [18] e o ID3 [19]. O algoritmo ID3 e o C4.5 são os mais populares.

2.1. Algoritmo de Indução C4.5 ou J4.8

Como formar uma árvore de decisão para um conjunto C de objetos? Se C é vazio ou contém somente objetos de uma mesma 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 $O_1, O_2, \ldots, O_w$. Existe um mapeamento entre cada objeto em C associado aos 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 resultado O_i. Se cada subconjunto C_i pode ser substituido por uma árvore de decisão, o resultado será uma árvore de decisão para todos os elementos de C. No pior caso essa estratégia fornecerá subconjuntos de um único objeto. Assim, uma vez que, um teste que gera uma divisão não trivial de qualquer conjunto de objetos sempre pode ser encontrado, este procedimento produzirá uma árvore de decisão que classifica corretamente os objetos em C [19].

A escolha do teste é crucial para a árvore de decisão ser simples. O algoritmo C4.5 adota um critério baseado na teoria da informação que depende de duas hipóteses:

- 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 que pertencem a classe P e n o número total de objetos pertencentes a classe N.

- Quando uma árvore de decisão é usada para classificar um objeto, ela retorna uma classe. Árvore de decisão pode ser considerada como uma fonte de mensagem P ou N em que a informação necessária para gerar a mensagem é obtida conforme 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 raíz da árvore de decisão, ela dividirá C em $\{C_1, C_2, \ldots, C_v\}$, 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 necessária para a subárvore 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, conforme equação (2.2)

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

em que o peso para o i-ésimo ramo é proporcional aos objetos em C que pertencem a C_i. Portanto, o ganho de informação obtido por esse ramo usando o atributo A é dado pela equação (2.3)

$$G(A) = I(p, n) - E(A). \tag{2.3}$$

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 [18, 19]. Na Figura 2 pode-se observar um fluxograma do algoritmo de construção de árvore de decisão.

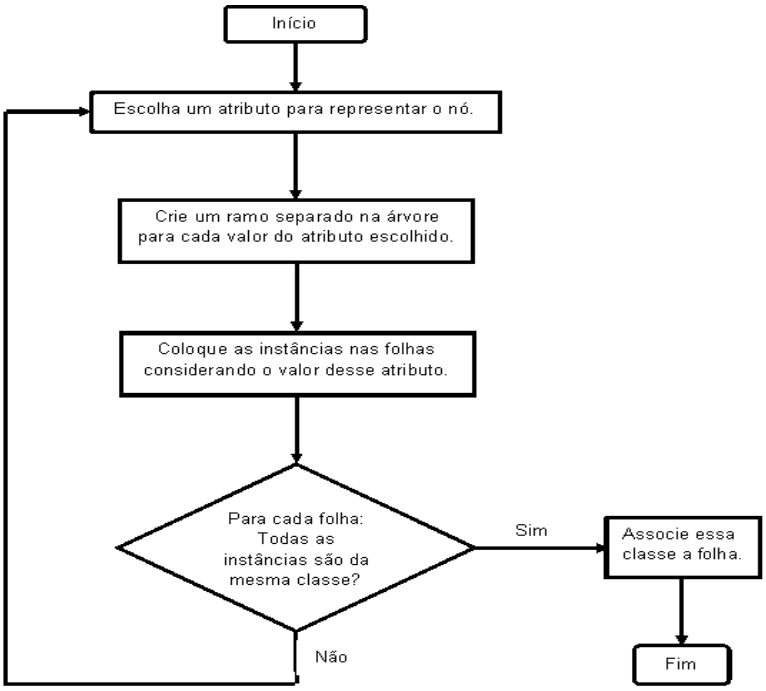

Figura 2: Fluxograma do algoritmo de árvore de decisão – Fonte: [18].

3. Aquisição dos Dados e Parâmetros Utilizados

Os dados utilizados neste trabalho são dados de seis anos do projeto SDSS [1]. 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* do SDSS, em todas as cinco bandas. O mesmo é feito para os dados espectroscópicos, de forma que há seis 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* é 1 no caso de estrela e 2 se galáxia. 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 mútiplos picos de intensidade nas imagens. A amostra final é composta por 43.289 estrelas e 452.400 galáxias. Os dados foram obtidos através do servidor CasJobs do SDSS [13, 23, 24].

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.

- **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 [21]. O parâmetro *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 [17], e 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 [21].

- **Redshift espectroscópico:** Este não é um dado fotométrico, mas sim obtido a partir dos espectros. Como é baseado em linhas de emissão e absorção, 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 o observado:

$$z = \frac{\lambda_{\text{Observado}} - \lambda_{\text{Emitido}}}{\lambda_{\text{Emitido}}} \tag{3.1}$$

  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.

4. Resultados Obtidos

O treinamento (criação da árvore) foi realizado com um conjunto de 10^4 objetos (925 estrelas e 9075 galáxias) utilizando-se o algoritmo J4.8. As 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 mas que correspondem a poucos casos no conjunto de dados fossem criadas. Na verdade, uma árvore de decisão é semelhante a um sitema especialista, pois fornece um conjunto de regras que devem ser aplicadas a um determinado objeto para obter sua classe. Nesta seção, será apresentado o desempenho de duas árvores que foram implementadas em linguagem C e testadas sobre o conjunto total de amostras (495.689 objetos).

O atributo *redshift* permite classificar de imediato os objetos em estrelas e galáxias. Assim, a amostra de treinamento continha todos os atributos fotométricos descritos na Seção 3, juntamente com o *redshift* espectroscópico. Nesse caso, o objetivo do treinamento era verificar a robustez do algoritmo J4.8 na identificação do atributo mais importante para a separação das classes. A árvore de decisão obtida com esse conjunto de treinamento definiu um valor crítico para o *redshift*: se *redshift* $\leq 0,001481$ o objeto é classificado como estrela, se *redshift* $> 0,001481$ o objeto é uma galáxia. A aplicação dessa regra sobre a amostra total forneceu um índice de acerto de 99.99%. Este resultado demonstrou a eficiência do algoritmo na escolha do atributo hegemônico na classificação.

Após este teste, o próximo passo foi remover o parâmetro *redshift* do conjunto de treinamento e realizar a criação de árvores somente com os outros parâmetros

fotométricos, portanto, cada objeto do conjunto de treinamento é descrito em termos de 40 parâmetros (8 atributos x 5 bandas). A estratégia de seleção de atributos utilizada pelo algoritmo J4.8 permitiu a identificação do parâmetro PetroR50 como o atributo que fornece uma melhor separação entre as classes, ou seja, que contêm *a maior quantidade de informação*. A Figura 3 exibe a primeira árvore e seu desempenho sobre o próprio conjunto de treinamento (10.000 objetos). Os números dentro dos retângulos que representam as classes (estrela ou galáxia) indicam a quantidade de objetos classificados corretamente/erroneamente por aquela folha. Para a criação desta árvore foi usado poda, fator de confiança de 0,25 e valor mínimo de 5 objetos por folha. Na Figura 4 tem-se a segunda árvore e também seu desempenho sobre o conjunto de treinamento. Os parâmetros de configuração foram os mesmos da primeira com exceção do número mínimo de objetos por folha que, nesse caso, foi estipulado 20 objetos.

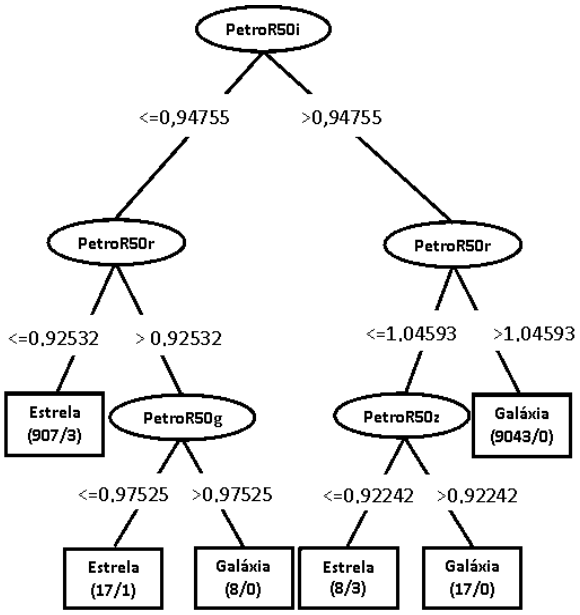

Figura 3: Esquema da primeira árvore de decisão

A Tabela 1 apresenta os resultados obtidos com as duas árvores sobre o conjunto de teste. Observa-se 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. 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.

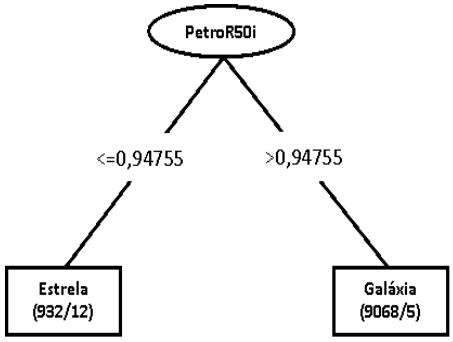

Figura 4: Esquema da segunda árvore de decisão

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%. Finalmente, na última linha da tabela tem-se o índice kappa, que é uma medida da acurácia da classificação, obtido por meio da matriz de confusão. Quanto mais próximo de 1 for o índice *kappa*, melhor é o desempenho do classificador.

Tabela 1: Resultados obtidos com as duas árvores de decisão sobre o conjunto de 495.689 objetos astronômicos.

	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%
Índice kappa	0,97		0,96	

Pode ser observado na Tabela 2 o índice de acertos referente a classificação de estrelas e galáxias do projeto SDSS obtidos com os classificadores desenvolvidos neste trabalho e também com os trabalhos de [22] e [3]. 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. No trabalho de [22] foi usado as diferenças $u - g$, $g - r$, $r - i$, $i - z$ e $g - i$. Já no trabalho [3] foi utilizado as mesmas cores que [22] 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 similar ao obtido nos trabalhos de [22] e [3]. O índice de acerto na classificação de estrelas foi cerca de 0,60% superior a [22] e cerca de 5, 30% superior a [3]. Para a classificação de galáxias o índice de acerto foi cerca de 1,00% superior a ambos os

trabalhos da literatura.

Tabela 2: Comparação entre os resultados obtidos neste trabalho e os obtidos pelos trabalhos de Suchkov et al. [22] e Ball et al. [3].

Classificadores	**Atributos utilizados**	**Índice de acerto**	
		Estrelas	**Galáxias**
Suchkov et al.[22]	*Cores dos objetos*	98,10%	98,50%
Ball et al.[3]	*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

A técnica de árvores de decisão foi empregada 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. 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 anteriores outras classes foram consideradas (e não somente estrela/galáxia); (b) há diferenças nos parâmetros fotométricos analisados e/ou limiares considerados; (c) estratégias da configuração das árvores de decisão (número mínimo de objetos por folhas, dentre outros); (d) os autores citados utilizaram outros algorítimos para implementar seus classificadores baseado em árvores de decisão.

Os classificadores desenvolvidos com árvores de decisão no presente trabalho alcançaram desempenho similar aos classificadores desenvolvidos por Suchkov et al. [22] e Ball et al. [3]. 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.

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.

Referências

[1] J. Adelman-Mccarthy et al., The sixth data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, **175**, No. 2 (2008), 297–313.

[2] N. M. Ball et al., Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks, *Monthly Notices of the Royal Astronomical Society*, **348** (2004), 1038–1046.

[3] N. M. Ball, R. J. Brunner, A. D. Myers, Robust machine learning applied to astronomical datasets I: star-galaxy classification of the sloan digital sky survey DR3 using decision trees. *The Astrophysical Journal*, **650** (2006), 497–509.

[4] D. Bazell, D. W. Aha, Ensembles of classifiers for morphological galaxy classification, *The Astrophysical Journal*, **548** (2001), 219–223.

[5] L. Breiman, Random forests, *Machine Learning*, **45**, No. 1 (2001), 5–32.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, "Classification and regression trees", U.S.A: Wadsworth Publishing Company, 1984.

[7] R.R Carvalho, H.V. Capelato, H.F. Campos Velho, Um universo escuro na era da tecnologia da informação, *Boletim da Sociedade Brasileira de Astronomia* - (submetido).

[8] F. Cortiglione, P. Mahonen, P. Hakala, T. Franti, Automated Star-Galaxy discrimination for large surveys, *The Astrophysical Journal*, **556** (2001), 937–943.

[9] Y. Freud, L. Mason, The alternating decision tree learning algorithm, *Proceedings of the Sixteenth International Conference on Machine Learning*, (1999), 124–133.

[10] J.P. Huchra, M.J. Geller, Groups of galaxies I. Nearby groups, *The Astrophysical Journal*, **257** (1982), 423–437.

[11] E.B. Hunt, J. Marin, P.J. Stone, "Experiments in Induction". New York: Academic Press, 1966.

[12] R. Kohavi, Scaling up the accuracy of naive - Bayes classifiers: a decision tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press*, (1996), 202–207.

[13] N. Lin, A.R. Thakar, CasJobs and MyDB: A batch query workbench, *Computing in Science and Engineering* , **10**, No. 1 (2008), 18–29.

[14] M.S. Madsen, "The Dynamic Cosmos - Exploring the Physical Evolution of the Universe", New York, NY, USA: Chapman e Hall, 1996.

[15] A.S. Miller, M.J. Coe, Star/galaxy classification using Kohonen self-organizing maps. *Monthly Notices of the Royal Astronomical Society*,**279**,(1996), 293–300.

[16] K.S. Murty, S. Kasif, S. Salzberg, A system for induction of oblique decision tree, *Journal of Artificial Intelligence Research*, **2**, (1994), 1–32.

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

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

[19] J.R. Quinlan, Induction of decision trees. *Machine Learning*, **1**, No. 1 (1986), 81–106.

[20] S. Salzberg et al., Decision trees for automated identification of cosmic ray hits in hubble space telescope images,*Publications of the Astronomical Society of the Pacific*, **107** (1995), 1–10.

[21] C. Stoughton, R.H. Lupton, M. Bernardi, M.R. Blanton, Sloan Digital Sky Survey: early data release. *The Astrophysical Journal*, **123**, (2002), 485–548.

[22] A. Suchkov, R.J. Hanisch, B. Margon, A Census of object types and redshift estimates in the SDSS photometric catalog from a trained decision tree classifier, *The Astronomical Journal*, **130**, (2005), 2439–2452.

[23] A.S. Szalay, A.R. Thakar, J. Gray, The sqlLoader data-loading pipeline, *Computing in Science and Engineering*, **10**, No. 1 (2008), 38–48.

[24] A.R Thakar, A.S. Szalay, G. Fekete, J. Gray, The catalog archive server database management system. *Computing in Science and Engineering*, **10**, No. 1 (2008), 30–37.

[25] I.H. Witten, E. Frank, "Data mining: Practical Machine Learning Tools and Techniques with JAVA Implementations". San Francisco: Morgan Kaufmann, 2000.

[26] Y. Zhang, Y. Zhao, A comparison of BBN, ADTree and MLP in separating quasars from large survey catalogues, *Chinese Journal of Astronomy and Astrophysics*, **7**, No. 2 (2007), 289–296.

[27] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, *Advances in Space Research*, **41**, No. 1 (2008), 1955–1959.