Um Processo Exploratório para Classificação de Estrelas e Galáxias

Eduardo Machado[1], Eduardo Bezerra[1,3], Ricardo Ogando[2,3], Marcio A. G. Maia[2,3], Luiz A. Nicolaci da Costa[2,3], Angelo Fausti Neto[3], Eduardo Ogasawara[1,3]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

[2]LINeA - Observatório Nacional

[3]Laboratório Interinstitucional de e-Astronomia (LIneA)

{ogando,ldacosta,maia,angelofausti}@linea.gov.br

ebezerra@cefet-rj.br, eogasawara@ieee.org

Resumo. *A separação entre estrelas e galáxias é fundamental para estudos galácticos e cosmológicos. Para fontes tênues, o limiar entre ser pontual e extenso fica confuso, dificultando o processo de classificação. O problema se agrava pelo grande volume de dados envolvido em levantamentos como o Dark Energy Survey (DES) e futuramente no Large Synoptic Survey Telescope (LSST). Nesse cenário, o LSST vai varrer o céu inteiro a cada três dias identificando milhões de fontes. Assim, a busca por métodos e processos que realizem a classificação com eficiência e acurácia é fundamental para que a comunidade de astronomia consiga escalar a sua capacidade de analisar os dados do LSST. Visando a preparação para esse contexto, este trabalho propõe a elaboração de um processo exploratório para classificação de estrelas e galáxias baseados na análise de catálogos treinados a partir do levantamento COSMOS.*

1. Introdução

Graças a telescópios de grande abertura e enorme mosaicos de detectores CCDs, levantamentos extensos e profundos do céu se tornaram uma realidade: vide SDSS, DES e, no futuro, o LSST. Um problema clássico da análise desses levantamentos consiste na classificação morfológica das fontes detectadas, separando fontes pontuais (ex. estrelas) de extensas (ex. galáxias). O problema começa quando fontes extensas muito distantes deixam de ser resolvidas podendo ser confundidas com fontes pontuais.

Comumente avaliam-se a qualidade da identificação de estrelas e galáxias por meio de duas medidas de qualidade: pureza e completeza. Um bom método de classificação deve obter valores elevados de completeza e pureza, esta última é particularmente importante em observações de seguimento (*follow-up*) de objetos especiais, de modo a não contaminar a amostra com falsas identificações, desperdiçando tempo de telescópio.

O presente trabalho tem por objetivo explorar os principais métodos de pré-processamento e de classificação para identificar se um objeto celeste corresponde a uma estrela ou a uma galáxia. A nossa abordagem inicial para o problema utiliza o levantamento COSMOS observado pelo Telescópio Espacial Hubble, o qual não é afetado pela turbulência (*seeing*) da atmosfera, dessa forma servindo de excelente amostra fiducial de

fontes pontuais e extensas. Nessa amostra empregamos métodos de pré-processamento de dados combinados a métodos de classificação.

2. Análise de Estrelas e Galáxias

A pesquisa relacionada ao problema da classificação estrela-galáxia usando métodos de mineração de dados ocorre há pelo menos duas décadas. Odewahn et al. [1992] descreveu um experimento utilizando redes neurais MPL com uma taxa de sucesso de 99% de classificação de galáxias com magnitude aparente M $<$ 18,5 e de 95% com 18,5 $<$ M $<$ 19,5. Posteriormente, diversos estudos foram realizados, dentre os quais destacam-se Digitized Sky Survey [Bazell and Peng, 1998], Sloan Digital Sky Survey (SDSS) [Elting et al., 2008] e Dark Energy Survey (DES) [Soumagnac et al., 2013].

As pesquisas realizadas no contexto de classificação estrela-galáxia podem ser organizadas pelos métodos de pré-processamento e classificação adotados. No que tange aos métodos de classificação, além das redes neurais [Qin et al., 2003], observou-se o uso de abordagens híbridas de redes neurais com lógica *fuzzy* [Mahonen and Frantti, 2000] e SVM [Elting et al., 2008].

No que se refere aos métodos de pré-processamento, observou-se a aplicação de funções de redução de dimensionalidade via *Principal Component Analysis* (PCA) [Soumagnac et al., 2013], remoção de outliers [O'Keefe et al., 2009] e técnicas de amostragem [Andreon et al., 2002]. O PCA foi usada para remover informação redundante ou insignificante de dados espectrais. Na amostragem, houve a preocupação de balancear a relação de estrelas e galáxias nos conjuntos de treinamento, uma vez que há muito mais estrelas que galáxias [O'Keefe et al., 2009].

Em linhas gerais, os estudos revelam que objetos brilhantes são classificados corretamente com maior acurácia. Para objetos de brilho mais fraco [Fadely et al., 2012], as técnicas apresentam erro significativamente maiores. Observou-se que quando métodos de pré-processamento não são empregados, a taxa de acerto na classificação pode ser severamente afetada [O'Keefe et al., 2009].

3. Processo Exploratório Proposto

O processo proposto contém sete atividades agrupadas em quatro etapas (particionamento, preprocessamento, método e avaliação) e está esquematizado na Figura 1. A etapa de particionamento (indicada em vermelho) contém a atividade de amostragem e consiste em separar o conjunto de dados dos catálogos em treinamento e teste. Na etapa de pré-processamento (indicada em verde), temos quatro atividades: remoção de valores extremos (*outliers*), normalização, seleção de atributos, transformação. Finalmente, as etapas de método (indicada em laranja) e avaliação apresentam, respectivamente, os métodos de classificação e de avaliação propriamente ditos. Neste processo, algumas atividades de pré-processamento são opcionais e podem ser suprimidas, como, por exemplo, a etapa de transformação. As atividades são descritas a seguir.

A atividade de amostragem consiste em separar o catálogo astronômico. As amostras podem ser produzidas de três formas gerais: aleatórias, estratificada e balanceada Lantz [2013]. A amostragem aleatória é a mais trivial, mas pode não ser a mais adequada. A amostragem estratificada procura manter a proporcionalidade do classificador existente

no conjunto de dados nas amostras de treinamento, enquanto que amostragem balanceada procura manter um equilíbrio do classificador nas amostras de treinamento.

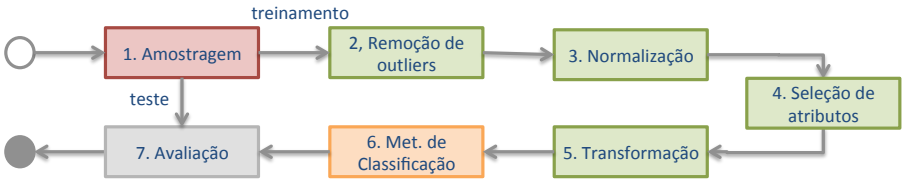

Figura 1. Processo exploratório proposto

A atividade de remoção de outliers contempla a análise de distribuição de dados e remoção de valores que destoam da distribuição dos dados. As opções consistem e considerar como outliers valores abaixo Q1-αIQR e acima de Q3+αIQR, em que IQR é o *interval quartil range,* e Q1 e Q3 representam o primeiro e terceiro quartis, respectivamente. O parâmetro α pode ser definido como 1,5 ou 3,0, dependendo da agressividade ou conservadorismo na percepção de outliers.

A atividade de normalização consiste em mapear os valores dos atributos de um conjunto de dados de tal forma a confiná-los a uma amplitude mais reduzida, tipicamente entre -1,0 e 1,0 ou 0,0 e 1,0. Assim, a normalização tenta dar a todos os atributos um peso igual, evitando que aqueles com grandes variações em seus valores se sobreponham aos de menor variação.

Os catálogos astronômicos contêm uma grande quantidade de atributos. Assim, a atividade de seleção de atributos consiste na redução da quantidade de atributos utilizados para elaboração do modelo, identificando os mais relevantes. Forward/backward stepwise selection e lasso James et al. [2013] são exemplos de métodos de redução de atributos. Essa redução visa evitar problemas de sobreajuste (*overfitting*).

Os métodos de transformação podem ser utilizados para redução de dimensionalidade. Um método bastante conhecido de transformação é o PCA, que transforma vetores de dados descritos por n dimensões que são projetados em um espaço de dimensão k ($k < n$), resultando em redução de dimensionalidade [Soumagnac et al., 2013].

Após a etapa de pré-processamento, pode-se aplicar diferentes métodos de classificação, como, por exemplo, Naïve Bayes, k-NN, Random Forests, Neural Networks e SVM [Han et al., 2011]. Os métodos podem ser utilizados individualmente ou podem ser combinados formando um modelo *ensemble*.

Finalmente, a etapa de avaliação consiste em aplicar o modelo de classificação previamente ajustado sobre o conjunto de testes para obtenção das medidas de pureza e de completeza.

4. Considerações Finais

Uma instanciação preliminar do processo proposto foi realizada no usando um sistema de workflow relacional [Ogasawara et al., 2013]. Todas as atividades previamente descritas foram implementadas como rotinas em R. Nessa instanciação preliminar, o workflow apresentou resultados onde as curvas ROC de classificação tiveram áreas acima de 0,925 e

a pureza se manteve acima de 97% para um valor de completeza de 96% e para magnitude i $<$ 23,5, o que demonstra que o processo proposto é promissor.

Referências

Andreon, S., Gargiulo, G., Longo, G., Tagliaferri, R., and Capuano, N. (2002). Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification: Wide field imaging - I. *Monthly Notices of the Royal Astronomical Society*, 319(3):700–716.

Bazell, D. and Peng, Y. (1998). A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination. *The Astrophysical Journal Supplement Series*, 116(1):47–55.

Elting, C., Bailer-Jones, C. A. L., Smith, K. W., and Bailer-Jones, C. A. (2008). Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines. pages 9–14. AIP.

Fadely, R., Hogg, D. W., and Willman, B. (2012). Star-Galaxy Classification in Multi-Band Optical Imaging. *The Astrophysical Journal*, 760(1):15.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, Mass., 3 edition edition. 24841.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer, 1st ed. 2013. corr. 4th printing 2014 edition edition. 00008.

Lantz, B. (2013). *Machine Learning with R*. Packt Publishing, Birmingham. 00019.

Mahonen, P. and Frantti, T. (2000). Fuzzy Classifier for Star-Galaxy Separation. *The Astrophysical Journal*, 541(1):261–263.

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., and Zumach, W. A. (1992). Automated star/galaxy discrimination with neural networks. *The Astronomical Journal*, 103:318.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., de Oliveira, D., Porto, F., Valduriez, P., and Mattoso, M. (2013). Chiron: a parallel engine for algebraic scientific workflows. *Concurrency and Computation: Practice and Experience*, 25(16):2327–2341. 00015.

O'Keefe, P. J., Gowanlock, M. G., McConnell, S. M., and Patton, D. (2009). Star-Galaxy Classification Using Data Mining Techniques with Considerations for Unbalanced Datasets. In *Astronomical Data Analysis Software and Systems XVIII*, volume 411, page 318.

Qin, D.-M., Guo, P., Hu, Z.-Y., and Zhao, Y.-H. (2003). Automated Separation of Stars and Normal Galaxies Based on Statistical Mixture Modeling with RBF Neural Networks. *Chinese Journal of Astronomy and Astrophysics*, 3(3):277–286.

Soumagnac, M. T., Abdalla, F. B., Lahav, O., Kirk, D., Sevilla, I., Bertin, E., Rowe, B. T. P., Annis, J., Busha, M. T., Da Costa, L. N., Frieman, J. A., Gaztanaga, E., Jarvis, M., Lin, H., Percival, W. J., Santiago, B. X., Sabiu, C. G., Wechsler, R. H., Wolz, L., and Yanny, B. (2013). Star/galaxy separation at faint magnitudes: Application to a simulated Dark Energy Survey. *arXiv:1306.5236 [astro-ph]*. arXiv: 1306.5236.