Combinação de Dados Tabulares e Imagens para a Classificação de Objetos Astronômicos

G. Jacob Perin*, L. Nakazono†, C. Mendes de Oliveira † e N. S. T. Hirata*
*Instituto de Matemática e Estatística
†Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Universidade de São Paulo

***Resumo*—No contexto do S-PLUS (*Southern Photometric Local Universe Survey*), um projeto de imageamento do céu do hemisfério sul em 12 bandas fotométricas, foram desenvolvidos anteriormente métodos baseados em aprendizado de máquina para a classificação de três tipos de objetos (estrelas, galáxias e quasares). Um dos trabalhos utilizou dados de catálogo, incluindo objetos com dados faltantes, enquanto o outro utilizou imagens, sem objetos com dados faltantes. Neste trabalho apresentamos dois avanços: avaliamos os dois métodos em condições de igualdade e propomos a utilização de técnicas de *ensemble* para combinar os dois tipos de classificadores. Experimentos realizados com o quarto *Data Release* do S-PLUS mostram que o *ensemble* proposto supera ambos os métodos anteriores com respeito aos objetos mais difíceis de serem classificados.**

***Abstract*—In the context of S-PLUS (*Southern Photometric Local Universe Survey*), a 12-band photometric sky survey of the southern hemisphere sky, two machine learning based methods were previously developed for the clasification of three types of objects (stars, galaxies and quasars). One of these works has used catalog data, including objects with missing information, while the other has used images, removing objects with missing information. In this work we present two advances: we evaluate the two methods under equal conditions and we propose the use of *ensemble* techniques to combine the two types of classifiers. Experiments performed with the fourth *Data Release* of the S-PLUS show that the proposed *ensemble* outperforms both previous methods with respect to the most difficult objects to be classified.**

I. Introdução

Uma característica fundamental da astronomia é o seu aspecto observacional. Para os astrônomos, os fótons capturados pelos telescópios possuem as principais informações que podem ser aprendidas sobre o universo. Nesse contexto, os recentes avanços tecnológicos relacionados ao imageamento do céu possibilitam a aquisição de enormes quantidades de dados. Dessa forma, surge a necessidade da criação de sistemas de software robustos e precisos para a análise desses dados.

O problema abordado nesse trabalho é a classificação de objetos astronômicos em 3 diferentes classes: estrelas, galáxias e quasares. No contexto do S-PLUS [1] (*Southern Photometric Local Universe Survey*), este problema foi abordado previamente em dois trabalhos, usando metodologias distintas. A principal diferença entre ambos é que, enquanto um utiliza *Random Forests* com dados tabulares e é robusto para informações faltantes [2], o outro utiliza redes neurais convolucionais no contexto de aprendizado auto-supervisionado e trabalha diretamente com as imagens de 12 canais do telescópio, porém utilizando uma seleção de dados mais conservadora e restritiva [3].

Neste trabalho, comparamos ambos os métodos propostos em condições de igualdade e propomos um *ensemble* que combina classificadores que utilizam dados tabulares e classificadores que utilizam imagens. Os resultados indicam que classificadores que utilizam dados de catálogo são superiores aos que utilizam imagens, e que o *ensemble* apresenta desempenho superior para os objetos mais difíceis de serem classificados do S-PLUS[1].

Nas seções a seguir, explicamos inicialmente alguns conceitos de astronomia. Em seguida, descrevemos a metodologia adotada (seleção dos dados e treinamento dos modelos) e os resultados obtidos. Por fim, apresentamos as conclusões.

II. Conceitos de Astronomia

Diversos tipos de objetos podem ser observados no céu. Nesse trabalho, três tipos de objetos são considerados para a classificação: **estrelas**, **galáxias** e **quasares**.

Nesse contexto, duas formas de se realizar medições desses objetos são chamadas de **espectroscopia** - que gera um espectro relacionando fluxo luminoso com comprimento de onda - e **fotometria**, que consiste em capturar imagens de comprimentos de onda específicos. Enquanto a primeira carrega mais informações e é mais demorada de ser realizada, a segunda é mais rápida, porém torna a classificação dos objetos mais desafiadora por carregar menos informação. O S-PLUS realiza um imageamento fotométrico, utilizando o sistema Javalambre de 12 filtros [1]. A figura 1 ilustra a fotometria de um quasar realizada pelo telescópio.

Para cada um dos 12 canais é possível calcular a sua **magnitude**: uma medida adimensional, de escala logarítmica, inversa ao brilho do objeto no respectivo canal. Nesse contexto, os dados tabulares utilizados pelo método proposto em [2] referem-se às medidas de magnitude dos objetos para cada um dos 12 filtros, além de 4 outras características morfológicas. Esses dados serão melhor explicados nas seções seguintes.

III. Metodologia

Os conjuntos de dados utilizados, assim como os modelos de classificadores, foram construídos usando os trabalhos anteriores como referência. No caso de dados tabulares, usamos

[1]Os códigos desenvolvidos para a análise dos dados e treinamento dos modelos podem ser encontrados, respectivamente, em https://github.com/gabjp/LTS2.0-data e https://github.com/gabjp/Label_The_Sky_2.0

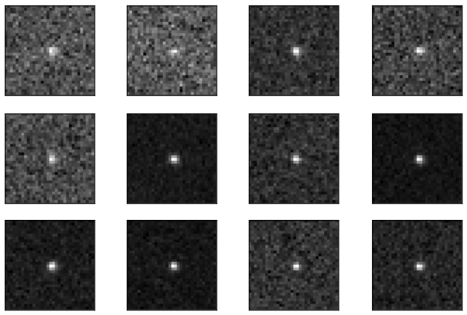

Figura 1. Imagem de 12 canais de um quasar. Quasares são extremamente luminosos e estão muito afastados da terra. Observa-se que, a olho nu, eles são fáceis de serem confundidos com estrelas

como referência o trabalho [2], e no caso das imagens usamos como referência o trabalho [3]. No entanto, algumas alterações foram realizadas para podermos fazer uma comparação direta entre os dois métodos. A seguir, descrevemos os conjuntos de dados e os modelos usados neste trabalho, apontando as alterações realizadas.

A. Conjunto de Dados

Os dados utilizados são provenientes de uma região do céu denominada *Stripe 82*, escolhida por já ter sido imageada diversas vezes pelo *Sloan Digital Sky Survey*[2] (SDSS) – um projeto de imageamento espectroscópico – e, portanto, uma ótima fonte para dados rotulados. Neste trabalho usamos o quarto *Data Release* (iDR4) do S-PLUS.

1) Dados tabulares: Os objetos das três classes de interesse são obtidos a partir do cruzamento dos dados do S-PLUS e do SDSS. A identificação da classe de um objeto é baseada no espectro disponível no SDSS.

Para remover dados excessivamente ruidosos, como em [2], foram retirados todos os objetos com magnitude R maior que 22 e estrelas com magnitude R menor que 13. Além disso, foram retirados objetos com marcação de erro pelo *SExtractor* (software de processamento das imagens utilizado na geração do catálogo). Ademais, foi aplicado um *sigma clipping* na diferença de magnitude R entre as medições do S-PLUS e do SDSS, para a remoção de *outliers*.

Desse processo resultaram 133128 objetos com rótulo de classe. Como um dos objetivos deste trabalho é a comparação dos dois métodos, procuramos uma divisão de dados que não prejudicasse nenhum dos métodos. Algumas formas de divisão foram testadas, e os melhores resultados foram obtidos com uma divisão análoga ao feito em [3]. Separamos 90% dos dados para treino, 5% para validação e 5% para teste (tabela I). A distribuição de magnitude R por classe do objeto e divisão é apresentada na figura 2.

Para o método que utiliza dados do catálogo, conforme [2], as seguintes *features* foram associadas a cada objeto: (i) as magnitudes das 12 bandas fotométricas do S-PLUS, presentes no catálogo do S-PLUS; (ii) as magnitudes de 2 bandas no

[2]https://www.sdss.org

Tabela I
CONJUNTO DE OBJETOS ROTULADOS

Classe	Treino	Validação	Teste	Total
Estrelas	49004	2722	2722	54448
Galáxias	54556	3028	3031	60615
Quasares	16262	903	900	18065
Total	119822	6653	6653	133128

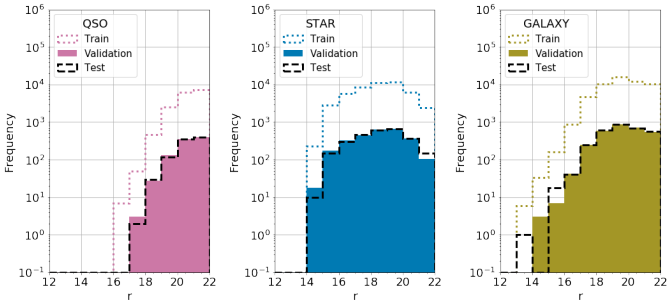

Figura 2. Distribuição da magnitude R por classe e divisão de dados. Observa-se que as distribuições são coerentes entre as divisões.

infravermelho do WISE (*Wide-field Infrared Survey Explorer*) [4]. WISE é um outro projeto de observação do céu cuja região observada tem intersecção com a região *Stripe 82*. Diferentemente do realizado em [2], para os objetos do S-PLUS que não estão presentes no WISE (cerca de 18%), atribuímos o valor 99 como magnitude dessas duas bandas (o mesmo valor usado nas bandas faltantes dos dados do S-PLUS); (iii) quatro características morfológicas: A, B, *Kron radius* e largura à meia altura (FWHM), presentes no catálogo do S-PLUS.

2) Imagens: Para o método baseado em imagens, recortes de 32×32 pixels foram feitos para cada um dos 133128 objetos selecionados acima. Caso o objeto fosse maior do que 32×32 pixels, utilizou-se o valor do FWHM como um representante para o seu tamanho. Realizou-se, então, o corte a partir desse valor e redimensionou-se a imagem para 32×32. Apenas cerca de 100 objetos precisaram ser redimensionados.

Em seguida aplicou-se uma calibração sobre cada imagem. A calibração é um processo que transforma a contagem de fótons em escala de brilho. No contexto de aprendizado de máquina, é um processo que facilita a convergência do modelo. Para o DR4 do S-PLUS, a calibração é dada pela seguinte expressão

$$x_{calib} = \frac{10^{5-0.4zp}}{ps^2}x \qquad (1)$$

na qual ps é a escala de pixel do detector (0.55 arcsec/pixel), zp é o valor de *zero point* (definido como a magnitude de um objeto que produz uma contagem por segundo no detector) e x é a imagem original. Observa-se que, enquanto ps é o mesmo valor para todas as imagens, zp varia com o campo e a banda utilizada.

Adicionalmente, criamos um conjunto de imagens da região *Stripe 82* sem rótulo de classe, para ser utilizado no pré-

treinamento de uma rede neural convolucional. Para tanto, amostramos cerca de 200 mil objetos usando os mesmos critérios de filtragem usados anteriormente e desde que os mesmos não estivessem no conjunto de objetos rotulados. As imagens dos objetos amostrados também passaram pelo mesmo processo de redimensionamento e calibração descritas anteriormente. Essa amostragem foi feita de forma que a distribuição das medições da magnitude R sejam semelhantes ao conjunto de objetos rotulados. Observa-se que, para que tal representatividade seja possível, diferentemente do que foi feito em [3], foram incluídos objetos com bandas faltantes (ou seja, valor 99). O conjunto de imagens sem rótulo de classe consiste de um total de 245261 objetos, sendo 211735 para treino, 11762 para validação e 11764 para teste.

B. Treinamento dos Modelos

1) Modelos com dados tabulares: Consideramos três classificadores *random forest* (com parâmetros `bootstrap=False` e `n_estimators=100`), que diferem entre si quanto à composição de dados usados no treinamento. Os dois primeiros são análogos aos classificadores descritos em [2], e o terceiro é uma alteração proposta neste trabalho.

- O primeiro classificador é baseado em todo o conjunto de dados de treinamento, sem utilizar as *features* do WISE (ou seja, utiliza as 12 bandas do S-PLUS e 4 *features* morfológicas). Denotamos esse classificador como **NW_RF** (*No-WISE_Random-Forest*).
- O segundo classificador é baseado apenas no conjunto de dados de treinamento com medições válidas do WISE (ou seja, com valores diferentes de 99) e utiliza as *features* do WISE. Denotamos esse classificador como **WW_RF** (*With-WISE_Random-Forest*).
- O terceiro classificador é baseado em todo o conjunto de dados de treinamento, e utiliza as *features* do WISE (nos objetos que não possuiam medições do WISE, essas *features* receberam valor 99). Denotamos esse classificador como **UF_RF** (*Unified_Random-Forest*).

Observamos que em [2], para a realização de predições, utiliza-se **WW_RF** caso o objeto tenha as duas bandas do WISE e utiliza-se **NW_RF** caso contrário. A motivação para o terceiro classificador é a unificação desses dois em apenas um, de forma a manter o bom desempenho de **WW_RF** proporcionado pelos dados do WISE e, ao mesmo tempo, utilizá-lo também para classificar objetos sem as medições no infravermelho.

2) Modelos com imagens: A arquitetura de rede neural usada é a VGG16 [5], implementada pela `tensorflow keras`, com mudanças nas camadas finais e iniciais para que a estrutura seja compatível com o problema.

A tarefa de pré-treinamento consiste em utilizar os dados não rotulados, especificamente as imagens das 12 bandas, e estimar os valores das 12 magnitudes extraídas do catálogo. A rede é treinada por 100 épocas, usando-se o otimizador *ADAM* com um *learning rate* de 10^{-5}. Diferentemente de [3], para que a rotina de treinamento possa comportar os valores faltantes nos *targets* (ou seja, valores de 99 nas magnitudes), usamos uma função de custo de erro absoluto médio modificada $L = \frac{1}{n}\sum e(y, \hat{y})$ na qual

$$e(y, \hat{y}) = \begin{cases} |y - \hat{y}|, & \text{se } y \neq 99 \\ 0, & \text{c.c.} \end{cases} \quad (2)$$

em que y são as magnitudes dos objetos, $\hat{y}$ é o valor de magnitude estimado pelo modelo e n é o número de previsões feitas (ou seja, 12 para cada objeto do conjunto de treinamento).

Na etapa de *fine-tuning*, a rede pre-treinada é treinada por 100 épocas usando as imagens do conjunto de dados rotulados. Diferentemente do proposto em [3], adotamos uma regularização l2 com parâmetro 0.0007 (escolhido a mão), para remediar o sobreajuste. Isso é esperado, visto que os dados rotulados utilizados no presente trabalho foram filtrados por critérios menos conservadores do que na tese da autora e, portanto, carregam uma maior quantidade de ruído que tende a agravar o sobreajuste. Esse classificador será denominado por **CNN**.

Tanto no pré-treinamento como no *fine-tuning*, no final de cada época os modelos são avaliados sobre o conjunto de validação e os pesos escolhidos são da época em que se obteve um melhor desempenho.

3) Modelos ensemble: Motivados pela ideia de que a grande diferença entre o espaço das *features* dos modelos baseados em *random forests* e dos baseados em redes convolucionais resultaria em uma grande diversidade de previsões, decidimos utilizar técnicas de *ensemble*.

Para a construção do *ensemble*, o *fine tuning* da rede pré-treinada é realizado usando-se apenas as imagens do conjunto de dados rotulados que não possuem medidas do WISE. Esse modelo é denominado por **SNW_CNN** (*Strict-no-WISE_CNN*).

No *ensemble* que propomos, a probabilidade de um objeto ser da classe t é modelada por $\alpha P^t_{SNW\_CNN} + (1-\alpha) P^t_{UF\_RF}$, $0 \leq \alpha \leq 1$, em que $P^t_{SNW\_CNN}$ e $P^t_{UF\_RF}$ são as probabilidades do objeto ser da classe t, de acordo com a **SNW_CNN** e o **UF_RF**, respectivamente. O parâmetro α é otimizado através de um *grid search* no conjunto de validação.

IV. Resultados

Cada um dos modelos foi treinado três vezes. No caso do **SNW_CNN** apenas a etapa de *fine-tuning* foi repetida 3 vezes. Assim foram gerados 3 classificadores para cada modelo. A avaliação de cada modelo é medida em termos de acurácia média e *F-score* médio. Observa-se que, como o nosso conjunto de dados não é balanceado entre as classes, a métrica de acurácia pode não refletir muito bem a performance do modelo e, portanto, o *F-score* médio representa uma medida mais adequada.

Como temos três classificadores **UF_RF** e três classificadores **SNW_CNN**, para o *ensemble* ordenamos os três de cada tipo em função do F-Score médio no conjunto de validação. Em seguida, pareamos as **SNW_CNN** e os **UF_RF** conforme a ordenação, e para cada par criamos um *ensemble*. Portanto, temos também três classificadores **Ensemble**.

Uma vez que os nossos dados não são representativos da proporção de objetos com e sem medidas do WISE (i.e., os objetos sem medidas do WISE são mais abundantes do que o nosso conjunto de dados sugere), dividiremos a nossa análise com respeito a essa distinção. A tabela II apresenta os resultados para o conjunto de validação com medidas do WISE, enquanto a tabela III apresenta os resultados para o conjunto de validação sem medidas do WISE.

Tabela II
RESULTADOS PARA O CONJUNTO DE VALIDAÇÃO COM MEDIDAS DO WISE

Experimento	Acurácia (%)	F-Score (%)
NW_RF	96.87 ±0.05	95.38 ±0.13
WW_RF	**98.44 ±0.03**	97.78 ±0.05
UF_RF	98.43 ±0.04	**97.79 ±0.08**
CNN	95.45 ±0.18	94.01 ±0.24

Na tabela II, observa-se que **WW_RF** e **UF_RF** possuem os melhores desempenhos, similares entre si. Portanto, o RF unificado proposto é melhor ou tão bom quanto os dois RF usados anteriormente. Ademais, comparando o desempenho do **NW_RF** entre as duas tabelas, destaca-se que os objetos que não possuem medidas do WISE são mais difíceis de serem classificados que os demais, mesmo quando a RF é treinada com todos os objetos (sem considerar as *features* do WISE). Isso ocorre pois o S-PLUS é mais profundo do que o WISE (i.e., consegue observar objetos de magnitude mais alta). Logo, é natural que os objetos que são detectados apenas pelo S-PLUS sejam menos brilhosos e, portanto, mais difíceis de serem classificados. Além disso, cabe observar nas duas tabelas que a **CNN** possui um desempenho notavelmente inferior aos modelos que trabalham diretamente com as magnitudes. Isso possivelmente ocorre pela dificuldade da CNN lidar com as incertezas associadas às imagens e magnitudes.

Para o conjunto de dados sem as *features* do WISE (tabela III), o desempenho superior do **SNW_CNN** sugere que a distribuição dos objetos que possuem e não possuem medidas do WISE é tão distinta que o ganho de restringir o domínio em **SNW_CNN** supera a perda devida à redução de dados. O **UF_RF** possui um desempenho superior ao **NW_RF**, o que sugere que a informação de que o objeto não foi detectado pelo WISE é útil para a classificação. Por fim, o **Ensemble** possui um desempenho superior a todos os outros modelos, demonstrando uma forma de aproveitar a diversidade obtida ao se trabalhar com representações diferentes dos objetos astronômicos.

Tabela III
RESULTADOS PARA O CONJUNTO DE VALIDAÇÃO SEM MEDIDAS DO WISE

Experimento	Acurácia (%)	F-Score (%)
NW_RF	85.94 ±0.18	85.29 ±0.18
UF_RF	87.91 ±0.47	86.87 ±0.49
CNN	81.44 ±0.66	80.89 ± 0.58
SNW_CNN	84.62 ±0.20	83.39 ±0.26
Ensemble	**88.48 ±0.35**	**87.52 ±0.39**

Tabela IV
RESULTADOS PARA O CONJUNTO DE TESTE SEM MEDIDAS DO WISE

Experimento	Acurácia (%)	F-Score (%)
NW_RF	83.36 ±0.07	83.15 ±0.06
UF_RF	85.87 ±0.07	85.28 ±0.08
Ensemble	**86.51 ±0.53**	**86.02 ±0.56**

Os resultados mostrados na tabela IV, referentes ao conjunto de teste, indicam que o fato de termos usado o conjunto de validação para ajustar o parâmetro α do **Ensemble**, não acarretou um sobreajuste ao conjunto de validação.

V. CONCLUSÃO

Reproduzimos os métodos de classificação de estrelas, galáxias e quasares anteriormente desenvolvidos no contexto do S-PLUS, e avaliamos ambos sobre um mesmo conjunto de objetos. Esta comparação direta mostrou que o classificador que utiliza dados do catálogo é superior àquele que utiliza imagens. Adicionalmente, tendo em mente que a geração dos catálogos de S-PLUS é feita utilizando o **NW_RF**, caso o objeto não possua medidas no WISE e o **WW_RF**, caso o objeto possua, propomos duas alternativas que exibem melhor desempenho para a classificação de objetos sem a medida do WISE: **UF_RF** (que possui desempenho superior e custo computacional semelhante ao método anterior) e o *ensemble* (que possui o melhor dos desempenhos, apesar do custo computacional adicional de se trabalhar com imagens e redes convolucionais). Cabe também observar que o **UF_RF** é um classificador unificado que dispensa a utilização seletiva de **WW_RF** ou de **NW_RF** (a depender de os objetos possuírem ou não medidas do WISE) para a classificação.

AGRADECIMENTOS

FAPESP, processos 2015/22308-2 e 2022/11645-1.

REFERÊNCIAS

[1] C. Mendes de Oliveira and *et.al.*, "The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies and redshifts with 12 optical filters," *Monthly Notices of the Royal Astronomical Society*, vol. 489, no. 1, p. 241–267, October 2019.

[2] L. Nakazono, C. Mendes de Oliveira, N. S. T. Hirata, S. Jeram, C. Queiroz, S. S. Eikenberry, A. H. Gonzalez, R. Abramo, R. Overzier, M. Espadoto, A. Martinazzo, L. Sampedro, F. R. Herpich, F. Almeida-Fernandes, A. Werle, C. E. Barbosa, L. Sodré Jr., E. V. Lima, M. L. Buzzo, A. Cortesi, K. Menéndez-Delmestre, S. Akras, A. Alvarez-Candal, A. R. Lopes, E. Telles, W. Schoenell, A. Kanaan, and T. Ribeiro, "On the discovery of stars, quasars, and galaxies in the Southern Hemisphere with S-PLUS DR2," *Monthly Notices of the Royal Astronomical Society*, vol. 507, no. 4, pp. 5847–5868, 07 2021.

[3] A. C. R. C. Martinazzo, "A self-supervised learning approach for astronomical images," Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo, October 2021.

[4] R. M. Cutri and *et al.*, "VizieR Online Data Catalog: AllWISE Data Release (Cutri+ 2013)," *VizieR Online Data Catalog*, p. II/328, Feb. 2021.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556