Tema
Tendências em Matemática Aplicada e Computacional, **18**, N. 1 (2017), 155-172

www.scielo.br/tema
doi: 10.5540/tema.2017.018.01.0155

Classificação Morfológica de Galáxias em Conjuntos de Dados Desbalanceados[†]

P. IANISHI e R. IZBICKI*

Recebido em 16 agosto, 2016 / Aceito em 31 março, 2017

RESUMO. Galáxias podem possuir diferentes morfologias, as quais são importantes fontes de informação para o entendimento da evolução do universo. O *Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey* (CANDELS) é um levantamento de milhares de imagens de galáxias distantes da Terra. Por não ser possível classificar todas essas imagens manualmente para descobrir suas respectivas morfologias, o desenvolvimento de classificadores automáticos precisos para tal tarefa é de extrema importância. Infelizmente, técnicas de predição tradicionais possuem baixo poder preditivo quando o conjunto de dados possui um forte desbalanceamento, ou seja, quando uma das classes da variável resposta é demasiadamente mais frequente do que as demais. Assim, este trabalho tem por objetivo estudar três abordagens que levam em conta a falta de balanceamento dos dados para o levantamento CANDELS e compará-los com os métodos usuais no problema de classificação de galáxias regulares e galáxias *merger*. Para comparar os diferentes métodos, diversas medidas de qualidade de métodos preditivos foram utilizadas. Mostramos que, para o caso de classificação de galáxias *merger*, as melhores predições foram provenientes das abordagens de sobreamostragem e mudança de corte. Para o caso de galáxias regulares, a importância de considerar o desbalanceamento foi menor, pois essa classe não possui um desbalanceamento tão forte quando comparada com a classe de galáxias *merger*. Além disso, mostramos que os classificadores obtidos via diferentes métodos de classificação (árvores de classificação, florestas aleatórias e regressão logística penalizada) levam a predições muito parecidas, o que indica que melhores predições só podem ser obtidas por meio da inclusão de novas estatísticas-resumo com base nas imagens ou por meio de bancos de dados maiores.

Palavras-chave: Classificação, conjunto de dados desbalanceados, aprendizado de máquina.

1 INTRODUÇÃO

Galáxias podem possuir diversas morfologias. Um esquema usual de classificação morfológica de galáxias foi criado por [9], que propõe as seguintes categorias principais (veja a Figura 1):

- **Galáxias Elípticas**. Têm uma distribuição suave de luz e têm a aparência de uma elipse.

[†]Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (2014/25302-2) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (200959/2010-7).
*Autor correspondente: Rafael Izbicki – E-mail: rafaelizbicki@gmail.com
Departamento de Estatística, UFSCar – Universidade Federal de São Carlos, 13565-905 São Carlos, SP, Brasil.
E-mail: paulaianishi@yahoo.com.br

- **Galáxias Espirais**. Consistem em um disco achatado, com estrelas formando uma estrutura espiral sobre ele.
- **Galáxias Irregulares**. Possuem uma morfologia perturbada e sem nenhum padrão. Possuem assimetria, núcleos descentralizados e estrutura irregular e caótica.

Além destas morfologias, galáxias também podem interagir umas com as outras. Em particular, quando elas estão se juntando são chamadas de galáxias *merger* (veja a Figura 2).

Tais categorias podem ser agrupadas em duas grandes classes: *regulares* – quando possuem forma espiral ou elíptica – e *não regulares* – quando são do tipo *merger*, de interação e/ou irregulares [8]. A Figura 3 apresenta exemplos de galáxias do conjunto investigado neste trabalho.

Figura 1: Exemplos de morfologia espiral, elíptica e irregular, respectivamente. Fonte: Wikimedia Commons.

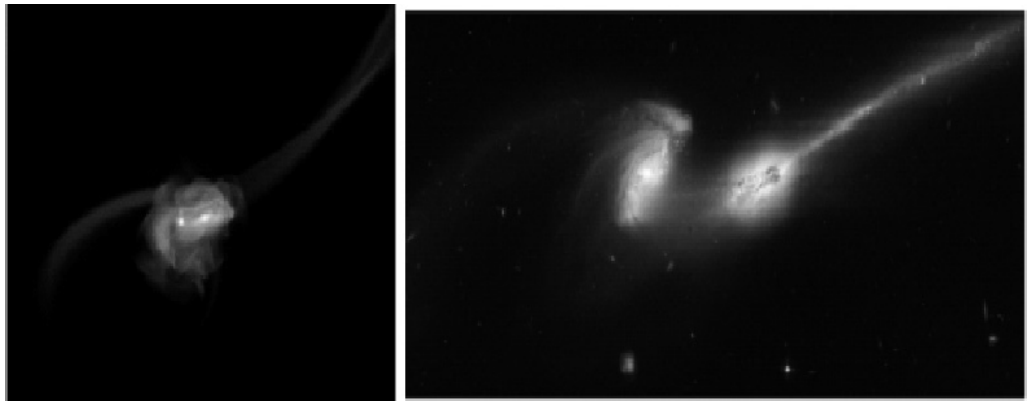

Figura 2: Exemplos de morfologia *merger* e de interação, respectivamente. Fonte: Wikimedia Commons.

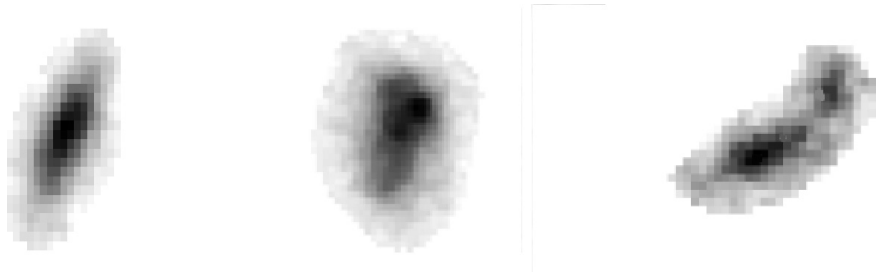

Figura 3: Exemplos de imagens do conjunto de dados CANDELS: galáxias elíptica, irregular e *merger*, respectivamente.

Estudar morfologicamente galáxias é fundamental para a corroboração de teorias sobre a formação e evolução cosmológica. Assim, é necessário um sistema que classifique eficientemente cada uma das imagens. Essa classificação pode ser feita por especialistas humanos, mas esse processo é escessivamente demorado tanto no seu desenvolvimento quanto em sua implementação [6]. Comumente, opta-se, portanto, pela utilização de classificadores automáticos, construídos com base em covariáveis extraídas de imagens [2, 6, 15].

Infelizmente, classificadores automáticos usuais não produzem resultados satisfatórios em situações nas quais uma das morfologias é demasiadamente mais frequente do que outras [21, 14, 20]. Essa situação ocorre frequentemente em levantamentos astronômicos como o *Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey* (CANDELS; [13]), conjunto este composto por 1639 galáxias e que foi utilizado no presente trabalho. Nele, observamos que aproximadamente 25% das galáxias são não regulares e que apenas 5% são *merger*. Assim, torna-se evidente a necessidade da aplicação de métodos mais sofisticados para tal problema. Neste trabalho, consideramos técnicas específicas para problemas de dados desbalanceados para que as classificações automáticas de galáxias sejam mais mais precisas. Além disso, utilizamos técnicas que não levam em conta o desbalanceamento, a fim de investigar a importância de considerar a falta de balanceamento para esse conjunto de dados. Embora diversos trabalhos utilizem correções para amostras desbalanceadas para a classificação morfológica de galáxias (e.g. [6, 16, 18]), os autores deste artigo desconhecem comparações entre as diferentes abordagens propostas na literatura como as feitas aqui.

Neste artigo, focamo-nos no desenvolvimento de métodos de classificação para galáxias do tipo não regulares e do tipo *merger*. Embora o foco deste trabalho seja o conjunto CANDELS, as técnicas aqui exploradas podem ser aplicadas a uma grande gama de problemas das mais diversas áreas do conhecimento.

O restante desse trabalho é dividido da seguinte maneira: a Seção 2 introduz os métodos de classificação utilizados neste artigo. Os resultados são apresentados na Seção 3. Finalmente, as conclusões são apresentadas na Seção 4.

2 METODOLOGIA

O conjunto de dados CANDELS é uma composição de fotografias de mais de 250 mil galáxias distantes da Terra feitas com três câmeras separadas no Telescópio Espacial Hubble [13]. Desse total de galáxias, 1639 foram classificadas manualmente por, pelo menos, dois astrônomos. Para que a classificação das galáxias fosse feita, os autores de [6] calcularam oito medidas resumo (chamadas C, S, M, I, D, A, Gini e m20) para cada uma das imagens do banco. Essas estatísticas são covariáveis a serem utilizadas pelo classificador automático e medem a concentração de luz, assimetria, presença de dois núcleos, além de outras características de cada uma das galáxias (vide Apêndice para mais detalhes). Assim, os dados são compostos de (I) classificação das galáxias segundo especialistas e (II) valores das estatísticas que ajudam a predizer a morfologia destas galáxias.

Pode haver discordâncias sobre a categoria à qual a galáxia pertence, a depender do especialista que a avalia. Assim, para definir a classificação de uma dada galáxia, utilizou-se o voto da maioria [4, 11]: uma galáxia foi classificada como não regular quando a proporção de votantes desta classe foi maior do que 50%. Do mesmo modo, uma galáxia foi classificada como *merger* quando a proporção de votantes desta classe foi maior do que 50%. Enfatizamos que, infelizmente, o conjunto de dados utilizado não possui informações sobre o voto de cada astrônomo. Contudo, em bancos nos quais esta identificação existe, modelos mais complexos que levam em conta a diversidade entre astrônomos podem ser utilizados (e.g. [11]).

Denotamos por $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ a amostra observada, em que $\mathbf{X}_i$ é o vetor de covariáveis C, S, M, I, D, A, Gini, m20 e Y é, em um primeiro momento, a variável que indica se a galáxia é regular ou não regular e, em um segundo momento, a variável que indica se a galáxia é ou não *merger* (i.e., foram resolvidos dois problemas de classificação separadamente)[1]. Assim, $Y \in \{0, 1\}$. De modo a comparar os diversos modelos ajustados, dividimos o conjunto de dados de forma aleatória em duas partes: um conjunto de treinamento (1230 amostras) e um de teste (409 amostras) [7, 12]. O conjunto de treinamento foi utilizado para ajustar diversas versões de um mesmo modelo (i.e., com diferentes parâmetros de *tuning*) e selecionar, entre elas, aquela que minimiza o risco estimado por validação cruzada. Já o conjunto de teste foi utilizado para avaliar o risco dos modelos selecionados. Todas as técnicas foram implementadas utilizando a linguagem R [19].

2.1 Métodos de classificação

Os seguintes métodos de classificação usuais (i.e., que não levam em conta o desbalanceamento dos dados) foram aplicados ao conjunto de dados [7, 12, 10]:

[1] As três categorias não são mutuamente exclusivas. Por exemplo, uma galáxia pode ser *merger* e não regular simultaneamente. Assim, o problema não pode ser trivialmente abordado sob uma ótica trinomial. Salientamos, também, que a ordem de solução dos problemas não influencia os resultados.

- **Árvores de classificação**. O critério de divisão utilizado foi o índice de Gini, que quantifica a pureza de uma dada folha[2] m via

$$\hat{p}_{m1}(1 - \hat{p}_{m1}),$$

em que $\hat{p}_{m1}$ é a proporção de amostras do conjunto de treinamento com rótulo 1 entre aquelas pertencentes à folha m. Tal índice foi escolhido pois (i) ele, em geral, é mais sensível à pureza que, por exemplo, a proporção de erros feita em cada folha [12] e (ii) é trivial adaptá-lo para o cenário com pesos. A profundidade da árvore foi escolhida por validação cruzada.

- **Florestas aleatórias**. Como recomendado por [12], o número de preditores considerados em cada divisão foi de $m \approx \sqrt{p}$, em que p é o número de covariáveis. Além disso, 500 árvores foram utilizadas.

- **Regressão logística penalizada**. Utilizou-se a penalização L1 [7], uma vez que ela faz, automaticamente, uma seleção de variáveis. O valor do parâmetro de penalização foi escolhido via validação cruzada. Estimado o valor de $P(Y = 1|\mathbf{x})$, tal quantidade foi substituída em $\mathbb{I}(P(Y = 1|\mathbf{x}) \geq 0.5)$ para criar um classificador.

Além dos métodos de classificação tradicionais descritos acima, foram também consideradas três abordagens para corrigir o desbalanceamento dos dados, descritas nas sequência.

Sobreamostragem. Esta abordagem consiste em criar artificialmente um conjunto de dados balanceado [20]. Isso foi feito acrescentando-se réplicas com reposição da amostra da categoria menos frequente até que os conjuntos de treinamento tivessem o mesmo número de observações em cada categoria. Os métodos descritos na Subseção 2.1 foram, em seguida, ajustados utilizando o conjunto de dados que fora balanceado. Nota-se que o conjunto de teste foi mantido, pois ele representa a população de interesse.

Atribuição de pesos. Nesta abordagem, atribuem-se pesos para cada observação. Em particular, atribuímos pesos maiores a observações de classes menos frequentes. Mais especificamente, o peso atribuído para a i-ésima observação foi:

$$w_i = \begin{cases} \frac{n_1}{n}, & \text{caso a classe dessa observação fosse a menos frequente} \\ \frac{n_2}{n}, & \text{caso contrário.} \end{cases}$$

Aqui, n_1 é o número de observações da classe mais frequente, n_2 é o número de observações da classe menos frequente e $n = n_1 + n_2$.

A forma como tais pesos são usados depende do método de classificação em questão. Para o caso do método de árvores, foi feita uma correção no índice de Gini, o qual, para uma dada folha m, passou a ser: $2\hat{p}'_m(1 - \hat{p}'_m)$, em que

$$\hat{p}'_m = \frac{\sum_{i \in folha(m): Y_i = 1} w_i}{\sum_{i \in folha(m)} w_i} = \frac{N_m \hat{p}_m \frac{n_1}{n}}{N_m \hat{p}_m \frac{n_1}{n} + N_m(1 - \hat{p}_m)\frac{n_2}{n}} = \frac{n_1 \hat{p}_m}{n_1 \hat{p}_m + n_2(1 - \hat{p}_m)},$$

[2]Isto é, é uma medida numérica de quão homogêneas são as categorias das observações referentes àquela folha.

em que N_m é o número de observações pertencentes à folha m e $\hat{p}_m$ é a proporção de observações na folha m que pertencem à classe 1. O mesmo procedimento foi utilizado na construção de cada árvore no caso de florestas aleatórias.

No caso da regressão logística penalizada, os pesos foram incluídos na função de verossimilhança. Assim, buscou-se pela solução de

$$\max_{\beta_0,\boldsymbol{\beta_1}} \left(\sum_{i=1}^{n} w_i [y_i(\beta_0 + \boldsymbol{\beta_1^T}\mathbf{x_i}) - \log(1 + e^{\beta_0 + \boldsymbol{\beta_1^T}\mathbf{x_i}})] - \lambda \sum_{j=1}^{p} |\beta_j| \right).$$

Aqui, $\boldsymbol{\beta_1}$ tem dimensão p, assim como $\mathbf{x_i}$.

Mudança do corte. O risco $R(g) = \mathbb{I}(g(\mathbf{X}) \neq Y)$, que motiva o uso dos classificadores tradicionais [7], não é adequado quando o conjunto de dados é desbalanceado. Por exemplo, para $g(\mathbf{X}) \equiv 0$, o risco da função $g(\mathbf{X})$ será baixo se $Y = 1$ ocorrer com frequência muito pequena, mas nenhuma nova observação será classificada como sendo da classe minoritária. Assim, como forma de contornar o problema, definiu-se uma segunda função de risco, dada por:

$$\begin{aligned} R'(g) &= \mathrm{E}[(\pi_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0)) + (\pi_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1))] = \\ &= \pi_1 \mathrm{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 0) + \pi_0 \mathrm{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 1), \end{aligned}$$

em que π_0 é a probabilidade de uma observação pertencer à classe $Y = 0$ e π_1 é a probabilidade de uma observação pertencer à classe $Y = 1$. Assim, dá-se maior importância ao erro de uma observação da classe 1 ser classificada como pertencente à classe 0 e menor importância ao erro de uma observação da classe 0 ser classificada como pertencente à classe 1. A função $g(\mathbf{x})$ que minimiza a esperança acima é dada por $g(\mathbf{x}) = \mathbb{I}(\mathrm{P}(Y = 1|\mathbf{x}) > \pi_1)$. De fato, a decisão ótima é $g(\mathbf{x}) = 1$ se, e somente se,

$$\pi_0 \mathrm{P}(Y = 0|\mathbf{x}) \geq \pi_1 \mathrm{P}(Y = 1|\mathbf{x}) \iff \mathrm{P}(Y = 1|\mathbf{x}) \geq \pi_1.$$

Isso motiva o uso do classificador $\mathbb{I}(\widehat{\mathrm{P}}(Y = 1|\mathbf{x}) \geq \widehat{\mathrm{P}}(Y = 1))$, em que $\widehat{\mathrm{P}}(Y = 1|\mathbf{x})$ foi estimada por meio dos métodos descritos anteriormente e $\widehat{\mathrm{P}}(Y = 1)$ é a proporção amostral da classe de interesse.

2.2 Qualidade do ajuste

Para avaliar a qualidade preditiva dos métodos investigados, as medidas utilizadas foram [20]: sensibilidade – $S = \frac{VP}{VP+FN}$; especificidade – $E = \frac{VN}{VN+FP}$; valor predito positivo – $VPP = \frac{VP}{VP+FP}$; valor predito negativo – $VPN = \frac{VN}{VN+FN}$; medida F – Medida F $= \frac{2}{\frac{1}{S}+\frac{1}{VPP}} = \frac{2.S.VPP}{S+VPP}$; média S. E. $= \frac{S+E}{2}$. Aqui, VP denota verdadeiro positivo; VN, verdadeiro negativo; FP, falso positivo e FN, falso negativo. Em um primeiro momento, positivo indica galáxia regular. Em um segundo momento, ser positivo indica galáxia *merger*.

3 RESULTADOS

Das 1639 galáxias no banco, 500 são não regulares (aproximadamente 30%) e 128 são do tipo *merger* (aproximadamente 8%). Na Subseção 3.1, apresentamos a performance de cada um dos métodos utilizados. Na Subseção 3.2, é feita a comparação entre os resultados via as estatísticas F e S.E., que sumarizam as demais. Já na Subseção 3.3, é avaliada a concordância entre as predições dos diversos métodos. Finalmente, na Subseção 3.4, ilustramos alguns dos classificadores obtidos.

3.1 Medidas de qualidade

As Tabelas 1 a 4 mostram as medidas de qualidade e seus respectivos intervalos de confiança 95% calculados por *bootstrap* [3] para os classificadores obtidos. Os resultados em negrito foram provenientes dos métodos que apresentaram as melhores medidas de qualidade em termos pontuais. Deve-se atentar que, em muitos dos casos, os intervalos de confiança indicam que, na realidade, estes valores são bastante parecidos com os demais.

Tabela 1: Medidas de qualidade para métodos usuais (i.e., sem correção por falta de balanceamento).

Regulares			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0.854	0.882	**0.918**
	(0.819, 0.889)	(0.847, 0.918)	(0.891, 0.945)
Especificidade	**0.453**	0.437	0.382
	(0.381, 0.526)	(0.364, 0.511)	(0.312, 0.454)
Valor predito positivo	0.774	**0.775**	0.765
	(0.735, 0.813)	(0.736, 0.814)	(0.727, 0.804)
Valor predito negativo	0.585	0.629	**0.680**
	(0.505, 0.667)	(0.554, 0.705)	(0.59, 0.771)

Merger			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0	**0.117**	0.088
	(0, 0)	(0.025, 0.21)	(0.007, 0.17)
Especificidade	1	**0.989**	0.986
	(1, 1)	(0.981, 0.998)	(0.977, 0.997)
Valor predito positivo	–*	**0.5**	0.375
	–*	(0.186, 0.814)	(0.07, 0.68)
Valor predito negativo	0.917	**0.925**	0.922
	(0.895, 0.939)	(0.904, 0.947)	(0.901, 0.945)

* valores que não puderam ser calculados, pois o divisor da fórmula foi zero.

3.2 Comparação entre os classificadores

A Figura 4 sumariza as principais medidas de qualidade para o problema de classificação de galáxias *merger*. Podemos observar que as correções nos métodos melhoram bastante as classificação de tais galáxias, o que é natural, visto que essa classe possui um desbalanceamento

Tabela 2: Medidas de qualidade por abordagem de sobreamostragem.

Regulares			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0.733	**0.846**	0.825
	(0.69, 0.776)	(0.812, 0.882)	(0.788, 0.863)
Especificidade	**0.640**	0.523	0.632
	(0.571, 0.711)	(0.45, 0.597)	(0.562, 0.704)
Valor predito positivo	0.817	0.795	**0.831**
	(0.777, 0.858)	(0.757, 0.835)	(0.794, 0.869)
Valor predito negativo	0.522	0.609	**0.623**
	(0.458, 0.587)	(0.534, 0.685)	(0.554, 0.692)

Merger			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	**0.617**	0.264	0.558
	(0.479, 0.756)	(0.138, 0.392)	(0.417, 0.7)
Especificidade	0.850	**0.970**	0.853
	(0.82, 0.881)	(0.956, 0.985)	(0.823, 0.884)
Valor predito positivo	0.272	**0.45**	0.256
	(0.189, 0.356)	(0.262, 0.638)	(0.173, 0.341)
Valor predito negativo	**0.960**	0.935	0.955
	(0.943, 0.978)	(0.915, 0.956)	(0.937, 0.974)

Tabela 3: Medidas de qualidade por abordagem de atribuição de pesos.

Regulares			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0.665	**0.882**	0.818
	(0.62, 0.711)	(0.851, 0.914)	(0.781, 0.856)
Especificidade	**0.664**	0.460	0.632
	(0.594, 0.734)	(0.388, 0.534)	(0.562, 0.704)
Valor predito positivo	0.813	**0.782**	**0.830**
	(0.769, 0.857)	(0.744, 0.821)	(0.793, 0.868)
Valor predito negativo	0.474	**0.641**	0.613
	(0.415, 0.535)	(0.559, 0.723)	(0.545, 0.683)

Merger			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	**0.647**	0.176	0.588
	(0.51, 0.784)	(0.066, 0.287)	(0.448, 0.729)
Especificidade	0.818	**0.986**	0.848
	(0.786, 0.852)	(0.977, 0.996)	(0.817, 0.879)
Valor predito positivo	0.244	**0.545**	0.259
	(0.17, 0.319)	(0.284, 0.807)	(0.177, 0.343)
Valor predito negativo	**0.962**	0.929	0.957
	(0.945, 0.98)	(0.909, 0.951)	(0.94, 0.976)

Tabela 4: Medidas de qualidade por abordagem de mudança de corte.

Regulares			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0.775	0.711	**0.829**
	(0.735, 0.816)	(0.668, 0.756)	(0.792, 0.866)
Especificidade	0.609	**0.703**	0.617
	(0.538, 0.68)	(0.637, 0.77)	(0.546, 0.689)
Valor predito positivo	0.813	**0.840**	0.826
	(0.774, 0.853)	(0.801, 0.88)	(0.789, 0.864)
Valor predito negativo	0.553	0.526	**0.622**
	(0.485, 0.621)	(0.465, 0.588)	(0.552, 0.692)

Merger			
	Árvores	**Florestas**	**Reg. Log. Pen.**
Sensibilidade	0	**0.705**	0.676
	(0, 0)	(0.576, 0.836)	(0.543, 0.81)
Especificidade	1	0.8	**0.853**
	(1, 1)	(0.766, 0.834)	(0.823, 0.884)
Valor predito positivo	-*	0.242	**0.294**
	-*	(0.171, 0.314)	(0.209, 0.38)
Valor predito negativo	0.917	**0.967**	**0.967**
	(0.895, 0.939)	(0.951, 0.984)	(0.951, 0.983)

* valores que não puderam ser calculados, pois o divisor da fórmula foi zero.

acentuado. Pode-se também notar que os métodos de sobreamostragem e pesos foram muito próximos em todos os casos. Isso ocorre porque os pesos aumentam artificialmente a importância de cada observação da classe menos frequente, do mesmo modo que a sobreamostragem aumenta o tamanho da classe menos frequente, igualando seu tamanho ao da classe mais frequente. Esta figura também indica que as árvores criadas segundo as abordagem de sobreamostragem e atribuição de pesos apresentam resultados superiores aos obtidos com a abordagem usual para o caso de galáxias *merger*. Além disso, árvores com correções de sobreamostragem e pesos apresentaram resultados melhores do que florestas, o que, em um primeiro momento, pode causar estranheza, pois árvores, em geral, possuem baixo poder preditivo. Isso pode ser justificado notando que o objetivo de florestas é diminuir o erro preditivo $E(\mathbb{I}(g(\mathbf{X}) \neq Y))$; observando as Tabelas 2 e 3, podemos concluir que isso realmente ocorreu. No entanto, florestas aumentaram o número de falsos negativos, o que diminuiu a sensibilidade, de modo que a performance de árvores foi melhor nesse sentido. Além disso, todas as galáxias foram preditas como não sendo do tipo *merger* para o caso de árvores sem correção e, portanto, mudar o corte não tem nenhum efeito: a probabilidade de uma observação pertencer à classe de galáxias que não são do tipo *merger* é estimada como 1. Devido a esse fato, o erro-padrão foi zero (Figura 4). Finalmente, observa-se que os resultados de florestas e regressão logística foram próximos quando utilizamos a abordagem de mudança de corte.

A Figura 5, que sumariza as principais medidas de qualidade para o problema de classificação de galáxias regulares, evidencia que as três abordagens que consideram o desbalanceamento

melhoraram o método de regressão logística para o caso de galáxias regulares. Este método foio melhor para predizer galáxias regulares. Em relação aos métodos sem correção, o melhor foi o de florestas aleatórias. Nota-se, contudo, que as correções nos métodos melhoraram muito mais os resultados da classificação de galáxias *merger* do que a de regulares, o que era esperado, visto que a primeira classe possui um desbalanceamento muito mais acentuado do que a classe de regulares.

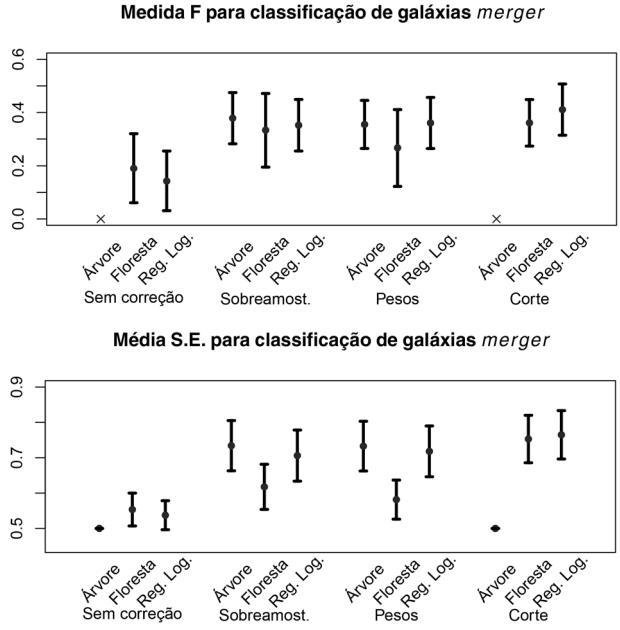

Figura 4: Medidas F (acima) e S.E. (abaixo) para classificação de galáxias do tipo *merger*.

3.3 Concordância entre os classificadores

Nesta subseção, investigamos o nível de concordância entre as predições fornecidas por cada um dos três modelos de predição utilizados na subseção anterior. Para tanto, selecionou-se, para cada abordagem, qual das quatro versões apresentava melhor performance segundo a estatística F. Os resultados para a estatística S.E. são semelhantes e, portanto, foram omitidos.

Para o caso de classificação de galáxias regulares, a melhor árvore de classificação e a melhor floresta aleatória foram aquelas com abordagem de sobreamostragem e a melhor regressão logística penalizada foi aquela com abordagem de mudança de corte. A Tabela 5 (esquerda) mostra a concordância dos melhores métodos. Para o caso de classificação de galáxias *merger*, a melhor árvore de classificação foi aquela sem correção (usual), a melhor floresta aleatória foi aquela com abordagem de atribuição de pesos e a melhor regressão logística penalizada foi aquela sem correção (usual). A Tabela 5 (direita) mostra a concordância dos melhores métodos.

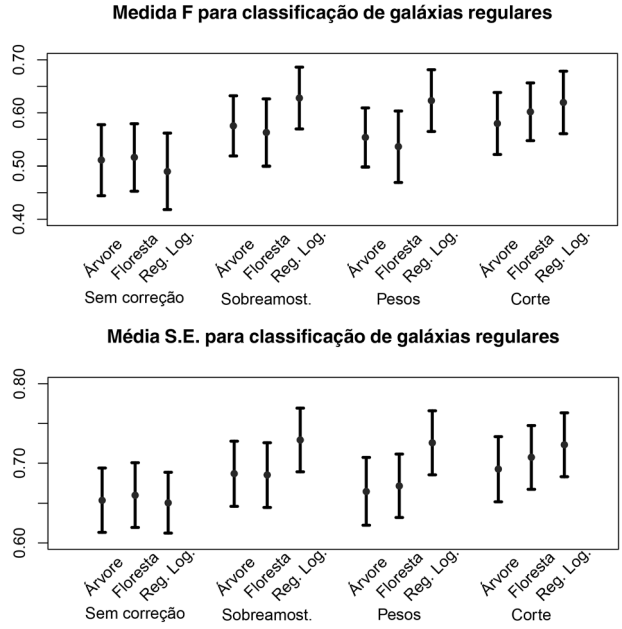

Figura 5: Medidas F (acima) e S.E. (abaixo) para classificação de galáxias regulares.

Tabela 5: Proporção de observações preditas igualmente por diferentes métodos – galáxias regulares (esquerda) e galáxias *merger* (direita).

	Árvore	**Floresta**	**R.L.P.**
Árvore	100.0%	91.4%	91.4%
Floresta	91.4%	100.0%	93.6%
R.L.P.	91.4%	93.6%	100.0%

	Árvore	**Floresta**	**R.L.P.**
Árvore	100.0%	86.1%	90.4%
Floresta	86.1%	100.0%	85.3%
R.L.P.	90.4%	85.3%	100.0%

A concordância é alta em ambas as tabelas, indicando que os melhores métodos levam a predições parecidas, apesar de terem naturezas bastante diferentes.

3.4 Ilustração dos classificadores obtidos

Em favor da concisão, apresentamos apenas os classificadores obtidos para classificação de galáxias *merger* para o método usual e para a abordagem de sobreamostragem, uma vez que os resultados omitidos levam a conclusões parecidas àquelas aqui apresentadas.

No caso da classificação pelo método de árvores, todas as galáxias do conjunto de teste foram classificadas como não sendo do tipo *merger*. Assim, sua representação gráfica foi omitida. De fato, 375 observações foram corretamente previstas como não sendo do tipo *merger* (100%), porém nenhuma observação foi corretamente predita como *merger*. Por outro lado, a Figura 6 evidencia que a árvore obtida considerando a abordagem de sobreamostragem é bem mais inte-

ressante. As estatísticas I, D e A são as covariáveis mais importantes segundo esta abordagem. Trezentas e dezenove (85%) observações foram corretamente previstas como não sendo do tipo *merger* e 21 (61.8%) observações foram corretamente previstas como sendo do tipo *merger*[3], o que também evidencia a melhora nas predições obtida ao se considerar o desbalanceamento.

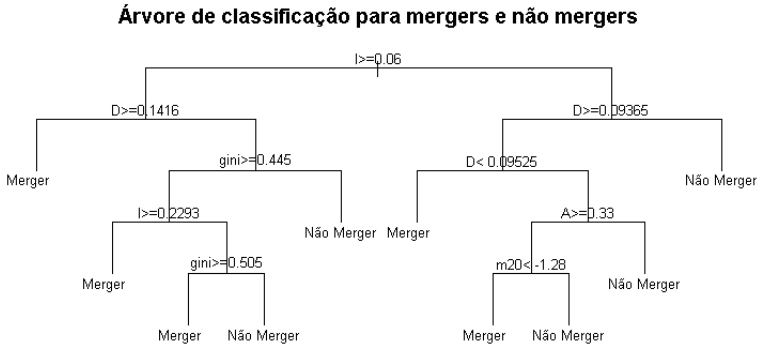

Figura 6: Árvore de classificação para galáxias *merger* obtida pelo método de sobreamostragem.

A Figura 7 indica que as covariáveis consideradas mais importantes para classificar galáxias *merger* segundo o método de florestas aleatórias sem correção foram D, I, e A. Trezentas e setenta e uma observações foram corretamente previstas como não sendo do tipo *merger* (aproximadamente 99%) e quatro observações foram corretamente preditas como *merger* (aproximadamente 12%). Quando corrigidas utilizando-se sobreamostragem, as covariáveis mais importantes na predição das galáxias *merger* segundo florestas aleatórias foram I, D, M e A, como indica a Figura 7. Além disso, nove observações foram corretamente previstas como não sendo do tipo *merger* e 364 observações foram corretamente previstas como sendo do tipo *merger*.

A Tabela 3.4 apresenta os coeficientes estimados segundo a regressão logística penalizada usual. As variáveis mais importantes na classificação de galáxias *merger* (i.e., variáveis associadas a coeficientes com maior magnitude) foram I, D e A. Além disso, 370 observações foram corretamente previstas como não sendo do tipo *merger* (aproximadamente 99%) e três observações foram corretamente preditas como sendo do tipo *merger* (aproximadamente 9%). A Tabela 3.4 também apresenta os coeficientes estimados para o mesmo método, mas com correção por sobreamostragem. As variáveis mais importantes foram, novamente, I, D e A. Além disso, 320 observações foram corretamente previstas como não sendo do tipo *merger* (aproximadamente 85%) e 19 observações foram corretamente preditas como sendo do tipo *merger* (aproximadamente 56%).

4 CONCLUSÕES

Neste trabalho, diversas técnicas de classificação foram aplicadas ao levantamento CANDELS com o objetivo de predizer automaticamente quais galáxias são não regulares e quais são do

[3]Isto é, 61.8% das galáxias *merger* do conjunto de teste foram classificadas como *merger*.

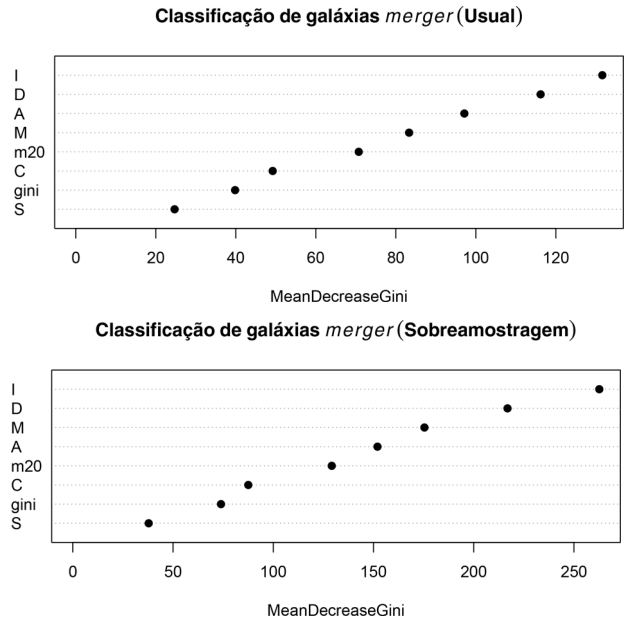

Figura 7: Classificação das galáxias regulares e não regulares pelo método de florestas aleatórias. Abordagem usual (à esquerda) e de sobreamostragem (à direita). O MeanDecreaseGini mede o quanto adicionar uma covariável na árvore diminui (em média) o índice de Gini [12].

Tabela 6: Coeficientes estimados pelo método de regressão logística penalizada usual (segunda coluna) e sobreamostragem (terceira coluna). Coeficientes estimados como zero por ambos os métodos são suprimidos.

Covariável	Usual	Sobreamostragem
(Intercepto)	–3.373	–1.872
M	0.001	0.000
I	1.993	2.761
D	0.051	1.826
A	2.928	4.942

tipo *merger*. Considerando que técnicas tradicionais apresentaram baixo poder preditivo por se tratar de dados desbalanceados, três correções a tais métodos foram utilizadas: sobreamostragem, atribuição de pesos e mudança de corte.

As medidas de qualidade de ajuste indicam que considerar o desbalanceamento não é tão importante para a classe de galáxias regulares. Isso ocorre pois essa classe não possui um forte desbalanceamento. Contudo, para o caso de galáxias *merger*, as abordagens que consideram o desbalanceamento melhoraram significativamente a performance dos classificadores usuais. Quando o desbalanceamento não foi considerado, o método árvore de classificação apresentou a pior performance. Em geral, com as devidas correções, este método apresentou grandes melhorias

nas medidas de qualidade, exceto no caso de mudança de corte. Isso ocorre pois árvores são construídas com o objetivo de minimizar a proporção de erros feita, e não de obter uma boa estimativa de $P(Y = 1|\mathbf{x})$. Por sua vez, o método de florestas aleatórias apresentou os melhores resultados quando o desbalanceamento não foi considerado. Porém, considerando as abordagens, ele é pior do que árvores para algumas situações. Finalmente, o método de regressão logística penalizada apresentou grandes melhorias quando consideramos as abordagens aqui estudadas, principalmente quando se muda o corte.

O fato de os métodos baseados em sobreamostragem terem resultados semelhantes a métodos com atribuição de pesos não é surpreendente. Deve-se destacar, contudo, que a vantagem do uso de pesos é que o tempo computacional para a sua execução é menor, pois não há necessidade de se trabalhar com um conjunto de dados maior. Por outro lado, a sobreamostragem é uma abordagem bastante geral que pode ser aplicada a qualquer classificador, ao passo que a forma com que os pesos são implementados é bem específica para cada método de classificação, o que faz com que a abordagem nem sempre seja trivial de ser implementada.

Semelhantemente ao que foi observado por [6], todos os métodos de classificação concordaram que as estatísticas I, D e A foram as mais importantes para a classificação tanto de galáxias regulares quanto de galáxias *merger*. Além disso, uma comparação entre valores preditos dos métodos que apresentaram melhor performance mostrou que os métodos levam a predições parecidas na maioria das vezes (concordâncias superiores a 85%). Isso indica que, possivelmente, melhores predições apenas podem ser obtidas por meio da inclusão de novas estatísticas-resumo com base nas imagens ou por meio de um banco de dados maior, e não pela aplicação de novos métodos a este banco. Alternativamente, pode-se buscar combinar os resultados dos métodos apresentados a partir de técnicas de *stacking*.

Outras direções futuras incluem: verificar a acurácia que cada um dos classificadores desenvolvidos tem ao estimar como cada morfologia evolui segundo *redshift* (tempo cósmico) (e.g. [1]), incorporar novas estatísticas resumo (inclusive algumas criadas automaticamente, e.g. [17]), utilizar técnicas multivariadas para classificar diversas morfologias simultaneamente [5] e, finalmente, utilizar técnicas semi-supervisionadas (i.e., que fazem uso de amostras não classificadas) para melhorar as predições [23].

A APÊNDICE – ESTATÍSTICAS USADAS PARA A CLASSIFICAÇÃO

Neste apêndice, descrevemos brevemente as estatísticas usadas para fazer a classificação automática. Mais detalhes podem ser encontrado em [6]. Denotamos por $f_{i,j}$ o valor do píxel (i, j) em uma dada imagem f em tons de cinza.

Estatística Multimode (M). Seja q_l um quantil de intensidade. Por exemplo, $q_{0,8}$ denota um valor de intensidade tal que 80 por cento das intensidades dos píxeis dentro do mapa de

segmentação são menores que esse valor. Inicialmente, com a finalidade de definir a estatística M, para um dado valor de l, considere uma nova imagem definida da seguinte maneira:

$$g_{i,j} = \begin{cases} 1, & \text{caso } f_{i,j} \neq q_l \\ 0, & \text{caso contrário} \end{cases}$$

Seja $A_{l,m}$ o número de píxeis em cada componente desta imagem, e seja

$$R_l = \frac{A_{l,(2)}}{A_{l,(1)}} A_{l,(2)},$$

em que $A_{l,(1)}$ é o maior grupo de píxeis adjacentes para o quantil l e $A_{l,(2)}$ é o segundo maior grupo de píxeis adjacentes. Essa estatística é utilizada para detectar a presença de dois núcleos no mapa de segmentação. Quando $\frac{A_{l,(2)}}{A_{l,(1)}}$ tende a 1, há presença de dois núcleos e, quando essa quantidade tende a 0, há a ausência. Como essa razão é sensível a ruídos, a multiplicamos por $A_{l,(2)}$, que tende a 0 caso o segundo maior grupo seja manifestação de ruído [6]. A estatística M é dada pelo máximo valor de R_l:

$$M = \max_l R_l$$

Estatística Intensidade (I). Inicialmente suavisa-se a imagem por meio de um kernel gaussiano bivariado simétrico [22]. Depois encontram-se os máximos locais utilizando o algoritmo *mean shift* (Figura 8).

Figura 8: Exemplo do tratamento da imagem, por agrupamento de píxeis, de uma galáxia *merger* para que seja possível computar a estatística I. As modas são encontradas pelo algoritmo *mean shift*.

A estatística I é então definida como

$$I = \frac{I_{(2)}}{I_{(1)}},$$

em que $I_{(1)}$ é a soma das intensidades dos píxeis ao redor de uma das modas e $I_{(2)}$ é a soma das intensidades dos píxeis ao redor da outra moda[4], com $I_{(1)} > I_{(2)}$.

[4]Aqui, "ao redor" é entendido como o quadrado de oito píxeis de altura e oito de largura centrado na moda.

Estatística Deviation (D). Seja (x_{cen}, y_{cen}) o centro de massa de uma imagem. A estatística D é definida como:

$$D = \sqrt{(x_{cen} - x_{I(1)})^2}\sqrt{(y_{cen} - y_{I(1)})^2},$$

em que $(x_{I_{(1)}}, y_{I_{(1)}})$ é o píxel onde a moda associada a $I_{(1)}$, definido anteriormente, se encontra.

Estatística A. A estatística A consiste na soma da diferença absoluta entre os píxeis da imagem original e da imagem rotacionada em 180°.

Estatística Concentração (C). A estatística C é definida como:

$$C = 5\log\left(\frac{r_{80}}{r_{20}}\right),$$

em que r_{80} e r_{20} são as aberturas circulares contendo 80% e 20% do fluxo total, respectivamente. A ideia é que se o raio da abertura que contém 80% for muito maior do que o raio que contém 20%, a razão $\frac{r_{80}}{r_{20}}$ será grande e isso é um indicativo de que há baixa concentração de luz. Por outro lado, se existe grande concentração de luz, os raios estarão muito próximos e a razão $\frac{r_{80}}{r_{20}}$ será próxima de 1, resultando em uma estatística C pequena.

Estatística (S). A estatística S é definida como

$$S = \sum_{i,j} \frac{|f_{i,j} - s_{i,j}|}{|f_{i,j}|} - B_S,$$

em que $s_{i,j}$ é a imagem suavizada e B_S é a suavidade média de fundo [15].

Estatística Gini. Seja $f_{(i)}$, com $i = 1, \ldots, n$, os fluxos ordenados dos píxeis, em que n é o número de píxeis na imagem. A estatística Gini é definida em [6] como:

$$\text{Gini} = \frac{1}{\bar{f}n(n+1)}\sum_{i=1}^{n}(2i - n - 1)f_{(i)}$$

sendo $\bar{f}$ a média de $f_{(i)}$. Para superfícies com luz pouco concentrada, essa estatística tende a zero e quando existe muita concentração de luz, a estatística tende a 1.

Estatística Momento de Luz (m20). A estatística m20 é uma medida de quão difusa a luz está na imagem e é definida como:

$$m20 = \log\left(\frac{\sum_{j \le i_{20\%}} f_{(j)}[(x_j - x_{cen})^2 + (y_j - y_{cen})^2]}{\sum_{j \in mask} f_{(j)}[(x_j - x_{cen})^2 + (y_j - y_{cen})^2]}\right),$$

em que $i_{20\%}$ é o valor de i para que $\sum_{j=1}^{i} f_{(j)} = 0.2\sum_{j=1}^{n} f_{(j)}$ e mask é a imagem original na qual se coloca um filtro para eliminar ruído.

AGRADECIMENTOS

Os autores agradecem a Adriano Polpo de Campos, Danilo Lourenço Lopes, Sarah Izbicki, os revisores e os editores pelas valiosas sugestões feitas a esse trabalho.

ABSTRACT. Galaxies can have various morphologies, which are an important source of information for cosmology. The *Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey* (CANDELS) is a survey of thousands of galaxy images far from the Earth. Unfortunately, it is not possible to manually classify all of these galaxies. Hence, it is important to develop automatic classifiers that are able to accurately predict morphologies using such images. Unfortunately, standard prediction techniques have low predictive power on unbalanced datasets such as CANDELS. Hence, this work aims at studying three classification approaches developed to improve classification on unbalanced data using CANDELS. We deal with the problem of classifying galaxies as regulars and as mergers. We show that oversampling and changing the cutoff were effective approaches to improve merger classification, while they were not so effective in classifying regular galaxies. We also show that all classification methods used (classification trees, random forests and penalized logistic regression) yielded similar predictions, which indicates that better predictions could only be obtained by including new summary statistics of the images or by acquiring larger data sets.

Keywords: Classification, unbalanced datasets, machine learning.

REFERÊNCIAS

[1] C.J. Conselice. The Evolution of Galaxy Structure Over Cosmic Time. *Annual Review of Astronomy and Astrophysics*, **52** (2014), 291–337.

[2] C.J. Conselice. The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, **147**(1) (2003), 1.

[3] B. Efron. The jackknife, the bootstrap and other resampling plans. **38** (1982), SIAM.

[4] L.G. Esteves, R. Izbicki & R.B. Stern. Teaching decision theory proof strategies using a crowdsourcing problem. *Submetido para American Statistician*, (2016).

[5] D. Fraix-Burnet, M. Thuillard & A.K. Chattopadhyay. Multivariate Approaches to Classification in Extragalactic Astronomy. In: *Frontiers in Astronomy and Space Sciences*, **2** (2015), 3.

[6] P.E. Freeman, R. Izbicki, A.B. Lee, J.A. Newman, C.J. Conselice, A.M. Koekemoer, J.M. Lotz & M. Mozena. New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, **434**(1) (2013), 282–295.

[7] J. Friedman, T. Hastie & R. Tibshirani. The elements of statistical learning. **1** (2001), Springer series in statistics Springer, Berlin.

[8] V.O. Gil, F. Ferrari & L. Emmendorfer. Investigação da aplicação de algoritmos de agrupamento para o problema astrofísico de classificação de galáxias. In: *Revista Brasileira de Computação Aplicada*, **7**(2) (2015), 52–61.

[9] E.P. Hubble. Extragalactic nebulae. In: *The Astrophysical Journal*, **64** (1926).

[10] R. Izbicki. *Machine Learning sob a ótica estatística*, (2016), `rizbicki.wordpress.com/teaching/`

[11] R. Izbicki & R.B. Stern. Learning with many experts: model selection and sparsity. *Statistical Analysis and Data Mining*, **6**(6) (2013), 565–577.

[12] G. James, D. Witten, T. Hastie & R. Tibshirani. *An introduction to statistical learning*. Springer (2013).

[13] A.M. Koekemoer, S.M. Faber, H.C. Ferguson, N.A. Grogin, D.D. Kocevski, D.C. Koo, K. Lai, J.M. Lotz, R.A. Lucas & E.J. McGrath et al. CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey – The Hubble Space Telescope Observations, Imaging Data Products, and Mosaics. *The Astrophysical Journal Supplement Series*, **197**(2) (2011), 36.

[14] S. Kotsiantis, D. Kanellopoulos & P. Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, **30**(1) (2006), 25–36.

[15] J.M. Lotz, J. Primack & P. Madau. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, **128**(1) (2004), 163.

[16] K. Małek, A. Solarz, A. Pollo, A. Fritz, B. Garilli, M. Scodeggio, A. Iovino, B.R. Granett, U. Abbas & C. Adami et al. The VIMOS Public Extragalactic Redshift Survey (VIPERS)-A support vector machine classification of galaxies, stars, and AGNs. *Astronomy & Astrophysics*, **557** (2013), A16.

[17] M.A. Peth, J.M. Lotz, P.E. Freeman, C. McPartland, S.A. Mortazavi & G.F. Snyder et al. Beyond spheroids and discs: classifications of CANDELS galaxy structure at $1.4< z< 2$ via principal component analysis. *Monthly Notices of the Royal Astronomical Society*, **458**(1) (2016), 963–987.

[18] M. Pović, J.A.L. Aguerri, I. Márquez, J. Masegosa, C. Husillos, A. Molino, D. Cristóbal-Hornillos, J. Perea, N. Benítez & A. del Olmo et al. The ALHAMBRA survey: reliable morphological catalogue of 22 051 early-and late-type galaxies. *Monthly Notices of the Royal Astronomical Society*, **435**(4) (2013), 3444–3461.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, (2016), `https://www.R-project.org/`

[20] Y. Sun, A.K.C. Wong & M.S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, **23**(04) (2009), 687–719.

[21] S. Visa & A. Ralescu. Issues in mining imbalanced data sets-a review paper. *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, **2005** (2005), 67–73.

[22] L. Wasserman. All of nonparametric statistics. Springer Science & Business Media, (2006).

[23] X. Zhu. Semi-supervised learning. Encyclopedia of machine learning. Springer, (2011), 892–897.