Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Hellen Mathei Della-Justina

Avaliação de Técnicas de Classificação Para Dados Desbalanceados

Curitiba
2023

Hellen Mathei Della-Justina

Avaliação de Técnicas de Classificação Para Dados Desbalanceados

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira

Curitiba
2023

Avaliação de Técnicas de Classificação Para Dados Desbalanceados

Evaluation of Classification Techniques for Imbalanced Data

Hellen Mathei Della-Justina[1]

[1]Aluna do Programa de Especialização em *Data Science* & Big Data, Departamento de Estatística, Universidade Federal do Paraná, Campus III - Centro Politécnico Rua Evaristo F. F. da Costa, 418, Curitiba, PR, Brasil *

Um conjunto de dados desbalanceado ocorre quando há diferença no número de amostras em diferentes classes. A fase de aprendizagem para a predição do modelo pode ser afetada em caso de dados desbalanceados. Então, neste estudo, foram aplicadas técnicas de *oversampling* e *undesampling* para lidar com dados desbalanceados. Os resultados mostraram um melhor desempenho do modelo *Random Forest* e das técnicas de *oversampling* para as métricas acurácia e precisão, um melhor desempenho das técnicas de *oversampling* para a métrica F1 e um melhor desempenho das técnicas de *undersampling* para as métricas *recall* e área sob a curva ROC.

Palavras-chave: dados desbalanceados, *random forest*, *oversampling*, *undersampling*

An imbalanced data occurs when there is a difference between the distribution of classes within a dataset. Machine learning models can be influenced by imbalanced datasets. So, in this study, it was applied the oversampling and undersampling techniques to deal with imbalanced data. The results show a better model performance for Random Forest and oversampling techniques for accuracy and precision metrics, a better oversampling performance for F1 metric, and a better undersampling performance for recall and ROC curve metrics.

Keywords: imbalanced data, random forest, oversampling, undersampling

1. Introdução

Um conjunto de dados desbalanceado ocorre quando há diferença no número de amostras em diferentes classes. Alguns conjuntos de dados são desbalanceados devido à própria natureza do dado. Por exemplo, dados de fraude em cartões de crédito, identificação de doenças raras, prevenção de *churn* dos clientes, classificação de spam em *e-mails*, entre muitos outros. Dessa maneira é muito importante haver modelos de aprendizagem de máquina que lidem com esse tipo de dados.

Os dados desbalanceados requerem um tratamento antes de alimentarem um modelo de aprendizado de máquina. Isso porque a fase de aprendizagem para a predição do modelo pode ser afetada em caso de dados desbalanceados [1]. A tendência é favorecer a classe com maior número de amostras, a chamada classe majoritária [2]. A performance de um modelo é mais gravemente afetada quando há conjuntos com classes desbalanceadas além de 90/10, ou seja, onde a classe minoritária apresenta menos de 10% das entradas totais.

Existem duas técnicas que são utilizadas para alterar a taxa das classes de amostra em um conjunto de dados desbalanceados: a *oversampling* e a *undersampling*. Enquanto a técnica de *oversampling* aumenta a quantidade de amostras da classe minoritária, a técnica de *undersampling* faz o contrário: ela diminui a quantidade de amostras da classe majoritária (Figura 1).

As três principais técnicas de *oversampling* são: *Naive Random Over Sampling*; *Synthetic Minority Over Sampling Technique* (SMOTE); e *Adaptive Synthetic*. A diferença entre estas técnicas é: enquanto a *Naive* duplica algumas amostras originais da classe minoritária a SMOTE e a *Adaptive Synthetic* (ADASYN) geram novas amostras pelo método de interpolação [3]. A interpolação é um método que permite construir um novo conjunto de dados a partir de um conjunto discreto de dados previamente conhecidos. A técnica SMOTE gera

*hellenjustina@gmail.com

novas amostras na área pertencente a regra do vizinho mais próximo, para maiores detalhes sobre a técnica SMOTE e ADASYN consulte [4] e [5], respectivamente.

As principais técnicas de *undersamplig* são: *Cluster Centroids*, a qual reduz as amostras da classe majoritária e gera uma nova amostra de dados. Diferentemente das técnicas *Random Under Sampling* e *Near Miss, Tomek's Links, Edited Nearest Neighbours, Repeated Edited Nearest Neighbours* e A11KNN, *Condensed Nearest Neighbour, One Side Selection, Neighbourhood Cleaning Rule, Instance Hardness Threshold* que selecionam as amostras do conjunto original [6], [7].

O objetivo principal desse projeto é: aplicar diferentes técnicas para classificar dados desbalanceados. Será feita a comparação entre métodos sem o desbalanceamento de classes e com a aplicação de diferentes métodos de desbalanceamento, pelas técnicas de *oversampling* e *undersampling*;

2. Materiais e Métodos

2.1. Programas e pacotes utilizados

A manipulação dos dados e a implementação das técnicas de dados desbalanceados foi realizada em linguagem de programação Python. Foram utilizados os pacotes *matplotlib, seaborn*, pandas e *numpy* para manipulação e visualização de dados; e o pacote *scikit-learn* para o modelo de aprendizado de máquina [8]. Ainda, utilizou-se a biblioteca *imbalanced learn*, sendo esta específica para a implementação das técnicas de classificação para dados desbalanceados [9].

2.2. Fonte dos dados

Como conjunto de dados foram utilizados os dados disponíveis publicamente pela biblioteca *imbalanced learn* [10]. Os dados são muito variados, como por exemplo: alguns conjuntos de dados médicos para classificação de doenças, um conjunto de dados que representam os crimes nos Estados Unidos, um conjunto de dados de petróleo, um conjunto de dados de níveis de ozonio, etc. Além disso, cada conjunto de dados apresenta uma proporção de desbalanceamento heterogênia.

2.3. Técnicas empregadas

Antes da implementação do modelo de aprendizado de máquina a base de dados foi dividida em treinamento e teste: 80% dos dados foi utilizado para treinamento e 20% para teste. É importante ressaltar que dessa forma os dados de teste continuaram desbalanceados. Em seguida, foi empregado o método *Random Forest* na base de dados original, ou seja, sem alteração na classificação das amostras desbalanceadas. Isto foi feito com o intuito de ter uma base de comparação, *baseline*, entre o desempenho das técnicas com um modelo padrão de classificação.

Em seguida, foram empregadas quatro técnicas de desbalanceamento, sendo duas técnicas de *oversampling*: *Random Over Samplig* e *SMOTE Over Samplig*; e duas técnicas de *undersampling*: *Random Under Sampling* e *Near Miss Under Sampling* (Figura 1). A avaliação das cinco diferentes técnicas foi empregada nos 27 *datasets* disponibilizados pela biblioteca *imbalanced learn*.

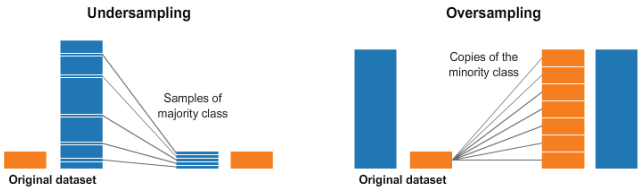

Figura 1: A técnica de *oversampling* aumenta a quantidade de amostrar da classe minoritária, enquanto que a técnica de *undersampling* faz o contrário: ela diminui a quantidade de amostras da classe majoritária.

Para aumentar a confiabilidade dos resultados, os cinco modelos foram repetidos trinta vezes em todos os 27 conjunto de dados. A comparação entre as métricas de classificação foi realizada pelo teste de Friedman [11], este é usado para avaliar se existem diferenças estatisticamente significantes entre distribuições de três ou mais grupos pareados.

2.4. Métricas de classificação

Uma questão extremamente importante é determinar qual métrica de classificação será utilizada para avaliação do desempenho das técnicas de desbalanceamento, e isto depende diretamente do tipo de dado a ser classificado. Por exemplo, quando se está interessado em detecção de fraude o importante é maximizar a taxa de verdadeiros positivos e capturar o máximo de fraudes possíveis, dessa forma o *recall*, ou sensibilidade, (expressa quão bem o modelo é capaz de detectar a classe correta) para a classe minoritária é a métrica que deve ser otimizada. Quando se trata da identificação de doenças raras o mais interessante é minimizar a taxa de falso positivo e não classificar erroneamente como comum qualquer doença consi-

derada rara. Dessa forma a precisão (expressa o quão confiável é o resultado quando o modelo responde que uma amostra pertence a classe correta) para a classe minoritária é a métrica mais indicada para otimização.

Outros fatores que indicam um alto desempenho de um modelo são altos valores para acurácia (número total de previsões corretas dividido pelo número total de previsões) e a área sob a curva ROC (esta mede a qualidade das previsões do modelo independentemente do limiar de classificação.Um modelo cujas previsões estão 100% erradas tem uma área sob a curva ROC igual a 0, enquanto um modelo cujas previsões são 100% corretas tem uma área sob a curva ROC igual a 1).

3. Resultados e Discussão

A Figura 2 ilustra um exemplo da distribuição das classes majoritária e minoritária de amostradas de um conjunto de dados utilizado para a avaliação das técnicas de *oversampling* e *undersampling*. A Figura 2 (b e c) ilustra um conjunto de dados após a implementação das diferentes técnicas de *oversampling*. A Figura 2 (d e e) ilustra o conjunto de dados após a implementação das diferentes técnicas de *undersampling*.

É possível observar a variação na quantidade de amostrar ajustadas nas diferentes técnicas. Houve um aumento do número de amostras da classe minoritária após a implementação das técnicas de *oversampling* e uma diminuição no número de amostras da classe majoritária após a implementação das técnicas de *undersampling*.

A avaliação das técnicas de *oversampling* e *undersampling* levou em consideração os parâmetros acurácia, precisão, *recall*, F1 e área sob a curva ROC como métricas de avaliação para o desempenho de classificação do modelo. Um alto valor da acurácia, da área sob a curva ROC e de *recall* associado a uma alta precisão significa que a classe é perfeitamente controlada pelo modelo.

A Figura 3 mostra a média das métricas utilizadas (acurácia, precisão, *recall*, F1 e área sob a curva ROC) após 30 execuções para cada uma das técnicas, com e sem desbalanceamento, nos 27 conjunto de dados.

Nos 27 conjuntos de dados testados os parâmetros de acurácia e precisão demostraram um melhor desempenho para o modelo *Random Forest* e para as técnicas de *oversampling* (*Random Over Sampling* e *SMOTE Over Sampling*). É possível observar, também, um melhor desempenho das técnicas de *oversampling* para o parâmetro F1. Além disso, os parâmetros com maiores valores para *recall* e área sob a curva ROC foram demonstrados após a aplicação das técnicas de *undersampling* (*Random Under Sampling* e *Random Under Sampling* usando *Near Miss*). De modo geral, considerando os valores dos parâmetros acima de 80%, o modelo que otimizou as cinco métricas utilizadas para realizar a comparação foi a técnica *SMOTE Over Sampling*.

Os resultados referentes ao teste de Friedman demonstraram uma diferença estatisticamente significante entre as métricas acurácia, precisão, *recall*, F1 e área sob a curva ROC (valores estatísticos acima de 60

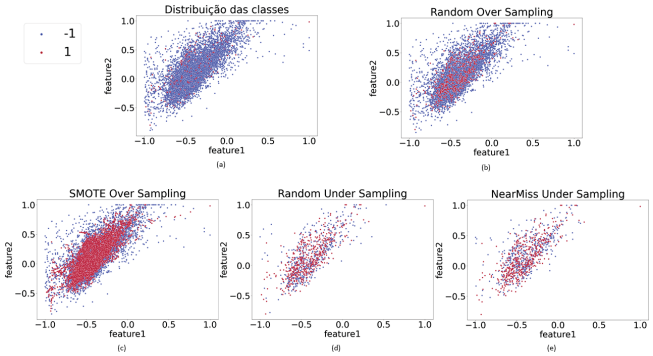

Figura 2: Distribuição das classes majoritária e minoritária. A *target* -1 (em azul) corresponde à classe majoritária e 1 (em vermelho) à classe minoritária. (a) Sem implementação de técnicas de desbalanceamento. (b) Após aplicar a técnica de desbalanceamento *Random Over Sampling*. (c) Após aplicar a técnica de desbalanceamento *SMOTE Over Sampling*. (d) Após aplicar a técnica de desbalanceamento *Random Under Sapling*. (e) Após aplicar a técnica de desbalanceamento *Random Under Sampling* usando *Near Miss*.

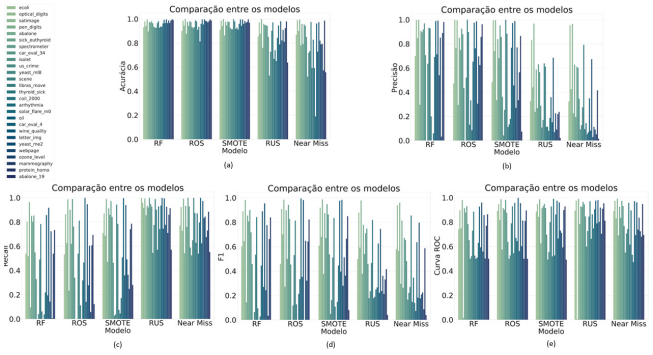

Figura 3: Média das métricas utilizadas para comparação dos cinco modelos empregados. No eixo x estão representados os modelos e no eixo y estão representadas as métricas empregadas. (a) Acurácia. (b) Precisão. (c) *Recall*. (d) F1. (e) Área sob a Curva ROC. RF - *Random Forest*; ROS - *Random Over Sampling*; SMOTE - *SMOTE Over Sampling*; RUS - *Random Under Sampling*; Near Miss - *Random Under Sampling* usando *Near Miss*.

e p-valores acima de 0,05) (Tabela 1). Um nível de significância de 0,05 indica que o risco de se concluir que existe uma diferença, quando, na verdade, não existe nenhuma diferença real, é de 5%. Ou seja, pode-se concluir que as métricas analisadas da população não são iguais.

Tabela 1: Resultados do teste de Friedman para os 27 conjunto de dados comparando as métricas para os cinco diferentes modelos.

Dataset	Métricas					Teste de Friedman
	Acurácia	Precisão	*Recall*	F1	Área sob a curva ROC	
ecoli	70.8028	88.1279	90.1623	77.2334	62.1933	valor estatístico
	1.5363e-14	3.28919e-18	1.2162e-18	6.7121e-16	1.0031e-12	p-valor
optical_digits	106.1149	104.9162	112.2013	95.0017	108.7893	valor estatístico
	4.9012e-22	8.8257e-22	2.4684e-23	1.1386e-19	1.3187e-22	p-valor
satimage	96.1286	111.7067	119.0317	103.6267	112.7467	valor estatístico
	6.5570e-20	3.1474e-23	8.5992e-25	1.6614e-21	1.8883e-23	p-valor
pen_digits	100.2065	102.1113	99.7000	92.4392	93.4733	valor estatístico
	8.8894e-21	3.4936e-21	1.1394e-20	3.9921e-19	2.4064e-19	p-valor
abalone	120.0000	70.3200	119.8063	111.2533	120.0000	valor estatístico
	5.3414e-25	1.9428e-14	5.8752e-25	3.9324e-23	5.3414e-25	p-valor
sick_euthyroid	90.9983	108.2915	103.2945	106.8476	58.8263	valor estatístico
	8.0802e-19	1.6838e-22	1.9554e-21	3.4208e-22	5.1178e-12	p-valor
spectrometer	90.7798	111.8062	76.1628	116.4674	64.3697	valor estatístico
	8.9917e-19	2.9972e-23	1.1309e-15	3.0338e-24	3.4932e-13	p-valor
car_eval_34	98.5616	95.2807	82.0531	107.6918	84.7200	valor estatístico
	1.9907e-20	9.9326e-20	6.3962e-17	2.2603e-22	1.7393e-17	p-valor
isolet	109.0569	113.8933	118.7605	113.5200	118.2437	valor estatístico
	1.5643e-22	1.0749e-23	9.8262e-25	1.2914e-23	1.2669e-24	p-valor
us_crime	97.8188	108.7346	112.6344	48.2609	109.0400	valor estatístico
	2.8648e-20	1.3547e-22	1.9954e-23	8.3269e-10	1.1660e-22	p-valor
yeast_ml8	104.9562	87.0830	114.7420	114.3860	108.6596	valor estatístico
	8.6540e-22	5.4818e-18	7.0839e-24	8.4386e-24	1.4054e-22	p-valor
scene	108.0268	80.1342	108.5347	109.7793	109.6267	valor estatístico
	1.9175e-22	1.6314e-16	1.4943e-22	8.1109e-23	8.7419e-23	p-valor
libras_move	113.0783	109.3992	96.6423	56.8190	65.1509	valor estatístico
	1.6044e-23	9.7751e-23	5.0983e-20	1.3502e-11	2.3916e-13	p-valor
thyroid_sick	82.1522	108.2428	89.6667	54.4482	103.8384	valor estatístico
	6.0939e-17	1.7246e-22	1.5499e-18	4.2398e-11	1.4975e-21	p-valor
coil_2000	119.0584	100.0267	114.3144	109.7333	113.6267	valor estatístico
	8.4870e-25	9.7088e-21	8.7407e-24	8.2958e-23	1.2255e-23	p-valor
arrhythmia	101.4879	60.0311	112.5295	119.0333	97.6765	valor estatístico
	4.7429e-21	2.8575e-12	2.1009e-23	8.5924e-25	3.0716e-20	p-valor
solar_flare_m0	119.0584	73.3600	120.0000	107.0400	101.0400	valor estatístico
	8.4870e-25	4.4278e-15	5.3414e-25	3.1125e-22	5.9077e-21	p-valor
oil	101.0629	86.7383	111.2628	100.0538	107.0400	valor estatístico
	5.8416e-21	6.4879e-18	3.91424e-23	9.5807e-21	3.1125e-22	p-valor
car_eval_4	102.6115	104.9462	82.0531	110.7273	46.3986	valor estatístico
	2.7336e-21	8.6969e-22	6.3961e-17	5.0918e-23	2.0345e-09	p-valor
wine_quality	118.5067	113.4357	115.6928	97.5200	114.4533	valor estatístico
	1.1132e-24	1.3460e-23	4.4395e-24	3.3165e-20	8.1638e-24	p-valor
letter_img	100.4682	98.3826	109.2068	117.2589	75.5726	valor estatístico
	7.8194e-21	2.1733e-20	1.0743e-22	2.0559e-24	1.5076e-15	p-valor
yeast_me2	108.4966	58.0203	116.6117	84.9008	105.2755	valor estatístico
	1.5225e-22	7.5565e-12	2.8261e-24	1.5922e-17	7.3993e-22	p-valor
webpage	114.2504	119.2267	120.0000	120.0000	117.2267	valor estatístico
	9.0198e-24	7.8131e-25	5.3414e-25	5.3414e-25	2.088e-24	p-valor
ozone_level	105.8003	106.2933	116.1109	112.5333	112.7813	valor estatístico
	5.7192e-22	4.4902e-22	3.6148e-24	2.0970e-23	1.8565e-23	p-valor
mammography	117.7229	115.7067	113.3605	100.3467	111.4133	valor estatístico
	1.6366e-24	4.4094e-24	1.3967e-23	8.2993e-21	3.6352e-23	p-valor
protein_homo	97.4637	118.5067	117.8400	112.0800	118.4533	valor estatístico
	3.4092e-20	1.1132e-24	1.5450e-24	2.6201e-23	1.1428e-24	p-valor
abalone_19	119.4573	110.8652	114.6319	113.7965	120.0000	valor estatístico
	6.9755e-25	4.7583e-23	7.4778e-24	1.1274e-23	5.3415e-25	p-valor

4. Conclusão

Neste trabalho foram empregadas técnicas de *oversampling* e *undersampling* para dados desbalanceados com a finalidade de otimizar a classificação por modelo de aprendizado de máquina. A utilização destas técnicas é extremamente importante antes da implementação de modelos de aprendizagem de máquina pois, dessa forma, a classificação das classes majoritária e minoritária serão melhor identificadas. Dentre os conjuntos de dados avaliados as técnicas de *undersampling* demonstraram melhores métricas para os parâmetros *recall* e área sob a curva ROC. Os resultados mostraram um melhor desempenho do modelo *Random Forest* e das técnicas de *oversampling* para as métricas acurácia e precisão e um melhor desempenho das técnicas de *oversampling* para a métrica F1.

Neste trabalho não foi realizada a avaliação entre a proporção de deslabanceamento do conjunto de dados e a otimização das métricas para os modelos empregadas na comparação. Dess forma, como trabalho futuro sugere-se realizar esse estudo com o intuido de identificar se a proporção de desbalanceamento entre a amostra sugere um melhor desempenho para as técnicas de *oversampling* ou *undersampling*. Além disso, o uso de simulação computacional para gerar o conjunto de dados desbalanceados poderia ser empregado com a finalidade de obter um maior controle dos resultados para identificação do modelo que otimize as métricas de comparação.

É importante destacar que as métricas de classificação devem respeitar a origem do conjunto de dados, cada conjunto deve ser avaliado de uma forma específica e interpretado seguindo os critérios característicos de sua origem para, assim, otimizar os seus parâmetros mais relevantes.

Agradecimentos

Gostaria de agradecer ao suporte dos professores do curso de especialização, em especial ao Prof. Dr. Wagner Hugo Bonat, pela dedicação e empenho em manter este curso e ao meu orientador, Prof. Dr. Luiz Eduardo Soares de Oliveira, pelo direcionamento do projeto. Agradeço também ao apoio incondicional de meus familiares, em especial ao meu esposo Lincoln e minha filha Sofia.

Referências

[1] B. Krawczyk, *Learning from imbalanced data: open challenges and future directions*, Progress in Artificial Intelligence, **5**, 221–232, 2016.

[2] G. Menardi and N. Torelli, *Training and assessing classification rules with imbalanced data*, Data Mining and Knowledge Discovery, **28**, 92–122 (2014).

[3] *Over-sampling methods*, **Imbalanced Learn**, 2023. Disponível em: `https://imbalanced-learn.org/stable/references/over_sampling.html`. Acesso em: 02 de fevereiro de 2023.

[4] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research, 321-357, 2002.

[5] H. He, Y. Bai, E. A. Garcia, and S. Li, *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning*, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008.

[6] *Under-sampling methods*, **Imbalanced Learn**, 2023. Disponível em: `https://imbalanced-learn.org/stable/references/under_sampling.html`. Acesso em: 02 de fevereiro de 2023.

[7] I. Mani and J. Zhang, *KNN approach to unbalanced data distributions: a case study involving information extraction*, Proceedings of Workshop on Learning from Imbalanced Datasets, **126**, 2003.

[8] *Scikit-learn: Machine Learning in Python*, **Scikit Learn**, 2023. Disponível em: `https://scikit-learn.org/stable/`. Acesso em: 01 de março de 2023.

[9] *Imbalanced-learn Documentation*, **Imbalanced Learn**, 2023. Disponível em: `https://imbalanced-learn.org/stable/index.html`. Acesso em: 01 de março de 2023.

[10] *Dataset Loading Utilities*, **Imbalanced Learn**, 2023. Disponível em: `https://imbalanced-learn.org/stable/datasets/index.html`. Acesso em: 01 de março de 2023.

[11] *Compute the Friedman test for repeated samples*, **SciPy documentation**, 2023. Disponível em: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.friedmanchisquare.html#r0143c258793d-2`. Acesso em: 28 de junho de 2023.