Lidando com o Desbalanceamento em Problemas de Classificação Hierárquica com Reamostragem de Dados

Rodolfo Miranda Pereira[1], Yandre M. G. Costa[2], Carlos N. Silla Jr.[1]

[1]Programa de Pós-Graduação em Informática (PPGIA)
Pontifícia Universidade Católica do Paraná (PUCPR) – Curitiba, PR – Brazil

[2]Departamento de Informática (DIN)
Universidade Estadual de Maringá (UEM) – Maringá, PR – Brazil

rodolfomp123@gmail.com, yandre@din.uem.br, carlos.silla@pucpr.br

Resumo. *Muitos problemas de classificação importantes são desbalanceados. Embora as abordagens de reamostragem sejam uma solução comum para diferentes tipos de problemas de classificação, elas ainda não foram definidas para problemas de classificação hierárquica. O objetivo deste trabalho é propor novas abordagens de reamostragem para lidar com a questão do desbalanceamento de classes em problemas de classificação hierárquica. Quatro direções foram investigadas: (i) O uso de métodos clássicos de reamostragem; (ii) Uma estratégia de conversão de caminho de rótulo; (iii) Esquemas para usar algoritmos de reamostragem com abordagens locais; (iv) Algoritmos de reamostragem globais. Para mostrar os impactos da contribuição deste trabalho, foi investigado o problema do desbalanceamento na identificação do COVID-19 em radiografias de tórax.*

1. Introdução

O desbalanceamento de classes é um problema em que o número de amostras de algumas classes é muito menor do que o número de instâncias de outras classes. Para lidar com este problema no contexto de classificação plana (binário, multiclasse e multirótulo), as técnicas de reamostragem (sobreamostragem e subamostragem) são as soluções mais bem-sucedidas.

Embora o desbalanceamento de classes seja um problema bem conhecido, poucos são os trabalhos que estudam essa questão no contexto da classificação hierárquica. Além disso, esses estudos não abordam diretamente o problema de desbalanceamento com métodos de reamostragem.

1.1 Objetivos

Os principais objetivos deste trabalho podem ser destacados como:

- Investigar a reamostragem binária/multirótulo em bases de dados hierárquicas.
- Propor métricas e técnicas para medir e lidar com o desbalanceamento em diferentes bases de dados hierárquicas, considerando as diferentes abordagens de classificação local e global.
- Investigar o uso das técnicas de reamostragem propostas em um estudo de caso real da sociedade.

1.2 Contribuições Científicas

A Tabela 1 apresenta um resumo dos artigos que foram enviados para publicação durante o desenvolvimento desta Tese. A tabela apresenta a referência do artigo (ou data de submissão), a relação do artigo e a (s) seção(ões) deste documento que descrevem

suas contribuições, o Impacto (Fator de Impacto (IF) para periódicos e o índice h5 (H5i) para conferências), o status da publicação e o número atual de citações.

Tabela 1. Artigos desenvolvidos durante esta pesquisa.

Referência	*Seções*	*Local*	*Impacto*	*Status*	*Citações**
[Pereira and Silla Jr. 2017]	2	ICME	H5i: 30	Publicado	3
[Valério et al. 2018]		FLAIRS	H5i: 16		3
[Pereira et al. 2019]		IJCNN	H5i: 46		1
[Mangolin et al. 2020]		Multimedia Tools and Application	IF: 2.31		1
[Pereira et al. 2020b]		Neurocomputing	IF: 4.44		8
[Pereira et al. 2018]	3 and 5	ICTAI	H5i: 19		3
[Pereira et al. 2020a]	2, 4, 5 and 6	Computer Methods and Programs in Biomedicine	IF: 3.63		138
[Pereira et al. 2021]	4	Data Mining and Knowledge Discovery	IF: 2.63		0
First submission on: February 26, 2020	5	Information Sciences	IF: 5.91	2ª Revisão	-
First submission on: March 19, 2021		Journal of Machine Learning Research	IF: 4.09	1ª Revisão	

*Citações obtidas do Google Scholar em 26 de maio de 2021.

1.3 Impacto Social e na Mídia

Este trabalho foi parcialmente desenvolvido durante o avanço da pandemia COVID-19. Considerando esse contexto, foi desenvolvido um método para identificar COVID-19 e outros patógenos de pneumonia em imagens de Raio-X de tórax (CXR). O trabalho foi um dos primeiros estudos publicados na literatura abordando essa questão [Pereira et al., 2020a]. Dada a importância do tema e a publicação oportuna dessa contribuição, ela tem atraído a atenção de pesquisadores, da sociedade e de veículos da mídia.

A fim de dar um panorama das repercussões, a Tabela 3 apresenta um resumo das principais reportagens sobre o trabalho na TV, no Rádio e em Revistas. Vale ressaltar que a reportagem da Agência Brasil foi republicada por mais de 100 sites de agências de notícias online, entre elas Valor Econômico, Época e Istoé.

Tabela 3. Resumo das principais reportagens da mídia sobre o trabalho.

Nome da Reportagem	*Nível de Repercussão*	*Local*	*Tipo de Mídia*
AI Assist for Spotting COVID-19 in X-Rays	Internacional	Physics	Revista
Coronavírus: Pesquisadores do Paraná criam método para diagnóstico com raio-x	Nacional	Gazeta do Povo	Revista
Estudo possibilita diagnóstico da COVID-19 via Raio-X	Nacional	CAPES	Youtube
Estudo brasileiro identifica COVID-19 com 90% de eficácia	Nacional	CNN Brasil	TV
Novo sistema de diagnóstico por Raio-X	Nacional	JovemPan News	Rádio
Pesquisadores do Paraná estudam método para diagnosticar COVID-19 com Raio-X	Nacional	CBN Maringá	Rádio
Radiografia Inteligente	Nacional	SuperAcesso	Revista
Universidades desenvolvem apoio a diagnóstico de COVID-19	Nacional	Agencia Brasil	Revista
COVID-19: estudo desenvolve diagnóstico por Raio-X	Estadual	CBN Curitiba	Rádio
Estudo com Inteligência Artificial para diagnóstico da COVID-19 em Raio-X	Estadual	Transamérica	Rádio
Imagens de Raio-X no diagnóstico da COVID-19	Estadual	RPC Paraná	TV
Pesquisa realizada por universidades paranaenses utilizam sistemas de inteligência artificial no diagnóstico da COVID-19	Estadual	Band News FM	Rádio

Pesquisadores fazem estudo p/ ajudar a identificar o COVID-19	Local	RIC Maringá	TV
Pesquisadores se unem para agilizar diagnóstico da COVID-19	Local	RPC Maringá	TV

2. O uso de reamostragem clássica em conjuntos de dados hierárquicos desequilibrados

Este foi o ponto de partida do trabalho, para entender os algoritmos de reamostragem existentes e como eles poderiam ser usados para lidar com os problemas de desbalanceamento em bases de dados de classificação hierárquica. Essas investigações foram capazes de responder às seguintes questões de pesquisa (RQs):

Q1.1: Pode-se aplicar os algoritmos de reamostragem clássica em bases de dados hierárquicas?

SIM: *Pode-se usar os algoritmos clássicos, no entanto, os caminhos de rótulo da base de dados devem ser considerados como rótulos individuais.*

Q1.2: Os algoritmos de reamostragem binários podem melhorar os resultados de classificação nas bases de dados hierárquicas com caminhos únicos?

PARCIALMENTE: *A análise experimental mostrou melhorias em apenas algumas bases de dados.*

Q1.3: Os algoritmos de reamostragem com vários rótulos podem melhorar os resultados da classificação nas bases de dados hierárquicas com múltiplos caminhos?

NÃO: A análise experimental não mostrou melhorias nos resultados da classificação.

3. Uma estratégia de conversão de caminho de rótulo

Esta foi a primeira proposta de esquemas específicos para lidar com a questão do desbalanceamento em bases de dados hierárquicas. As principais contribuições são duas métricas para medir o desequilíbrio (IRLP e HMeanIR) e dois algoritmos de conversão (HMC → ML e ML → HMC). Os algoritmos HMC↔ML são capazes de converter os caminhos de rótulo em formatos de rótulo múltiplo, portanto, pode-se aplicar algoritmos de reamostragem de multirótulo nas bases de dados hierárquicas. Considerando este tópico, pode-se investigar as seguintes RQs:

Q2.1: Pode-se desenvolver um método de conversão de bases de dados hierárquicas em bases de dados multirótulo sem perder informações de relacionamento dos rótulos?

SIM: *Pode-se usar a árvore de rótulos da base de dados para converter os rótulos em um formato multirótulo e, em seguida, usá-lo novamente para convertê-lo de volta.*

Q2.2: Pode-se medir o desbalanceamento de uma base de dados hierárquica de uma forma global?

SIM: *Com a proposta do IRLP e do HMeanIR, pode-se determinar o nível de desbalanceamento de um determinado conjunto de dados hierárquicos.*

Q2.3: A estratégia de conversão proposta pode aumentar os resultados da classificação?

SIM: *A estratégia foi estatisticamente capaz de melhorar os resultados.*

4. Abordagens de reamostragem local

Dentre as técnicas para lidar com a classificação hierárquica, as abordagens locais, ou seja, Classificadores Locais por Nós (LCN), Classificadores Locais por Nó Pai (LCPN) e Classificadores Locais por Nível (LCL) são abordagens bem conhecidas na literatura.

Nesse contexto, este trabalho propõe: (i) três novas métricas (IRLCN, IRLCPN e IRLCL) para medir o desbalanceamento em conjuntos de dados de classificação hierárquica considerando as três diferentes abordagens locais; e (ii) novos esquemas de classificação para lidar com o desbalanceamento em conjuntos de dados hierárquicos usando os algoritmos de reamostragem plana nas abordagens locais. Considerando este tópico, pode-se investigar os seguintes RQs:

Q3.1: Pode-se medir o desbalanceamento nos conjuntos de dados hierárquicos considerando as abordagens LCN, LCPN e LCL?

SIM: *Usando as amostras que resultaram nos subconjuntos locais gerados durante as etapas de treinamento em cada perspectiva local diferente, podemos calcular uma relação quantitativa entre as amostras rotuladas com cada nó.*

Q3.2: Os algoritmos de reamostragem simples podem melhorar os resultados nas abordagens LCN, LCPN e LCL?

SIM: *Os esquemas propostos melhoraram estatisticamente os resultados.*

Q3.3: Os esquemas de reamostragem local propostos podem reduzir o desbalanceamento considerando as métricas propostas (IRLCN, IRLCPN e IRLCL)?

SIM: *O Teste de Correlação de Person foi capaz de identificar uma correlação entre os resultados da classificação e as métricas propostas.*

Q3.4: A política usada para selecionar as amostras de subconjunto durante as etapas de treinamento local de LCN e LCPN influencia na reamostragem?

SIM: *As diferentes políticas são consideradas nos esquemas de reamostragem ao construir os subconjuntos locais.*

5. Abordagens de reamostragem global

Um algoritmo de reamostragem global deve ser capaz de lidar com a hierarquia de rótulos em um conjunto de dados como um todo, lidando com os diferentes tipos de problemas de classificação hierárquica, como a profundidade da predição e o número de caminhos associados a cada amostra. Foram propostos três novos algoritmos: (i) Superamostragem hierárquica aleatória (HROS); (ii) Subamostragem hierárquica aleatória (HRUS); e (iii) Técnica Hierárquica de Sobreamostragem Sintética (HSMOTE). Esses algoritmos foram capazes de responder as seguintes RQs:

Q4.1: Pode-se definir uma maneira de recuperar os conjuntos majoritários e minoritários de caminhos de rótulos em um conjunto de dados hierárquico?

SIM: *Usando o IRLP e o HMeanIR, pode-se recuperar o conjunto majoritário/minoritário de caminhos de rótulo, selecionando os caminhos de rótulo com um IRLP abaixo/acima do HMeanIR.*

Q4.2: Pode-se lidar com os diferentes tipos de problemas hierárquicos?

SIM: *Ao lidar com os problemas de profundidade parcial, a reamostragem deve ser feita percorrendo-se a árvore de rótulos de forma ascendente, recalculando o desbalanceamento para atualizar os rótulos maioritários/minoritários.*

Q4.3: Pode-se produzir conjuntos sintéticos de caminhos de rótulo combinando instâncias de vizinhos?

SIM: *Diferentes combinações foram desenvolvidas considerando os tipos de problemas hierárquicos: caminhos únicos ou múltiplos com profundidade total ou parcial de*

predição. Nas combinações de profundidade parcial, os subcaminhos devem ser considerados durante as combinações.

6. O estudo de caso de identificação COVID-19 em imagens raio-x do tórax

O principal caso de estudo deste trabalho foi a identificação de diferentes tipos de pneumonia causadas por múltiplos patógenos usando características texturais de imagens de raio-x do tórax. Especificamente, foram consideradas a pneumonia causada por vírus (COVID-19, SARS, MERS e Varicela), bactérias (Streptococcus) e fungos (Pneumocystis). Esses patógenos (rótulos) foram organizados hierarquicamente de acordo com suas relações biológicas. Foi desenvolvida uma base de dados inovadora para a investigação, pela qual pôde-se responder às seguintes RQs:

Q5.1: Pode-se usar recursos de textura das imagens de raio-x do tórax para reconhecer patógenos de pneumonia?
SIM: *Foram testados nove descritores de textura diferentes nas imagens e alguns deles mostraram resultados promissores para a tarefa investigada.*

Q5.2: As técnicas computacionais de reconhecimento de padrão podem diferenciar os tipos de patógenos causadores da pneumonia?
PARCIALMENTE: *Na análise experimental, alguns patógenos obtiveram uma taxa de reconhecimento promissora, como o SARS-CoV-2 e MERS, contudo, outros patógenos não tiveram taxas satisfatórias, como a Varicela e o Pneumocystis.*

Q5.3: Um esquema de classificação hierárquica pode ter um desempenho melhor do que um esquema de classificação plano para a tarefa de identificação COVID-19?
SIM: *Os melhores resultados de classificação foram alcançados usando as abordagens hierárquicas com reamostragem, propostas neste trabalho.*

Q5.4: Os esquemas de reamostragem propostos podem melhorar os resultados da classificação sem abordagens de reamostragem?
SIM: *Dentre as propostas, os melhores resultados foram alcançados usando a classificação hierárquica local LCN com o esquema de reamostragem proposto.*

Q5.5: A proposta está pronta para ser implementada nas unidades de saúde?
PARCIALMENTE: *O trabalho mostra que existe sim a possibilidade de profissionais de saúde serem auxiliados na detecção de patógenos de pneumonia em imagens. Contudo, deve-se investigar bases de dados maiores para uma maior confiabilidade.*

7. Observações Finais

Neste trabalho, foram propostas novas abordagens para lidar com a questão do desbalanceamento para diferentes tipos de problemas de classificação hierárquica. As duas métricas globais propostas foram capazes de avaliar o quão desbalanceado é uma base de dados hierárquica (Q2.2 e Q4.1).

Foi proposto um método baseado na conversão do conjunto de dados hierárquico em um formato multirótulo, a fim de aplicar algoritmos de reamostragem multirótulo bem conhecidos no conjunto de dados para lidar com o desequilíbrio. A abordagem foi capaz de melhorar os resultados quando comparados a aplicação da abordagem de reamostragem plana diretamente nos conjuntos de dados hierárquicos (Q2.1 e Q2.3).

No principal estudo de caso deste trabalho, os esquemas de reamostragem propostos para lidar com bases de dados hierárquicas desbalanceadas usando as abordagens de classificação local alcançaram os melhores resultados se comparados com todas as outras abordagens propostas (Q3.1, Q3,2, Q3.3, Q3.4, Q5.3 e Q5.4).

Também foram propostos três novos algoritmos de reamostragem global capazes de lidar com conjuntos de dados hierárquicos como um todo (Q4.2 e Q4.3). Em geral, todos os algoritmos também foram capazes de melhorar estatisticamente os resultados em determinados cenários.

Por meio deste trabalho, foi possível identificar uma importante contribuição para um problema do mundo real que afeta atualmente nossa sociedade: A detecção de patógenos de pneumonia, como o SARS-COV-2, em imagens de raio-x do tórax. Primeiramente, foi identificado que esse problema também pode ser enquadrado como um problema de classificação hierárquica, uma vez que existem relações biológicas entre os patógenos causadores da pneumonia (Q5.2). Em segundo lugar, trata-se de um problema de classificação que sofre de grave desbalanceamento (Q5.3), visto que existe um desbalanceamento natural entre o número de pessoas com pulmões saudáveis e o número de pessoas com pulmões afetados por pneumonias causadas por determinados tipos de patógenos, como como COVID-19.

Vale ressaltar que este trabalho deu origem a projetos nacionais e internacionais de conclusão de curso, dissertações de mestrado e teses de doutorado, principalmente relacionados ao tema de detecção de COVID-19 em imagens (Q5.5).

Referências

Mangolin, R. B., Pereira, R. M., Britto Jr., A. S., Silla Jr, C. N., Feltrim, V. D., Gonçalves, D. B., Costa, Y. M. G. (2020) A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, pages 1-26.

Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla Jr., C. N., Costa, Y. M. G. (2020a). COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, 194(C):1-18.

Pereira, R. M., Costa, Y. M. G., Aguiar, R. L., Britto, A. S., Oliveira, L. E. S., Silla Jr., C. N. (2019). Representation Learning vs. Handcrafted Features for Music Genre Classification. In: *Intl. Joint Conference on Neural Networks*, pages 1-8.

Pereira, R. M., Costa, Y. M. G., Silla Jr, C. N. (2018). Dealing with Imbalanceness in Hierarchical Multi-Label Datasets Using Multi-Label Resampling Techniques. In: *Intl. Conference on Tools with Artificial Intelligence*, pages 818-826.

Pereira, R. M., Costa, Y. M. G., Silla Jr, C. N. (2020b). MLTL: A multi-label approach for the Tomek Link undersampling algorithm. *Neurocomputing*, 383(C): 95-105.

Pereira, R. M., Costa, Y. M. G., Silla Jr, C. N. (2021). Handling imbalance in hierarchical classification problems using local classifiers approaches. *Data Mining and Knowledge Discovery*, 1-58.

Pereira, R. M.; Silla Jr, C. N. (2017). Using simplified chords sequences to classify songs genres. In: *Intl Conference on Multimedia and Expo*, pages 1446-1451.

Valério, V. D., Pereira, R. M., Costa, Y. M. G., Gonçalves, D. B., Silla Jr, C. N. (2018). A Resampling Approach for Imbalanceness on Music Genre Classification using Spectrograms. In: *Florida A.I. Research Society Conf.,* pages 500-505.