Dealing with Imbalanceness in Hierarchical Classification Problems Through Data Resampling

Rodolfo M. Pereira[1]*, Yandre M. G. Costa† and Carlos N. Silla Jr.*
*Programa de Pós-Graduação em Informática (PPGIA)
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, Paraná, Brazil 80215–901
†Departamento de Informática (DIN)
Universidade Estadual de Maringá (UEM)
Maringá, Paraná, Brazil 87020–290

***Abstract*—Many important classification problems are imbalanced. Although resampling approaches are a common solution for different types of classification problems, they were still not defined for hierarchical classification problems. The objective of this work is to propose novel resampling approaches to handle the class imbalanceness issue in hierarchical classification problems. Four directions were investigated: (i) The use of classic resampling methods; (ii) A label path conversion strategy; (iii) The design of schemas to use resampling algorithms with local approaches; (iv) The proposal of global resampling algorithms. To show the impacts of the contribution of this work, we have investigated the imbalanceness issue in the COVID-19 identification in chest x-ray images.**

***Keywords*- hierarchical classification; class imbalance; resampling algorithms**

I. Introduction

Class imbalance is an issue where the number of samples from some classes is far less than the number of instances from other classes. In order to deal with this problem in the flat classification context (binary, multi-class and multi-label), the resampling techniques (oversampling and undersampling) are the most successful solutions [1].

Although class imbalance is a well-known problem, there are few works studying this issue in the context of hierarchical classification. Furthermore, these studies do not directly address the imbalance problem with resampling methods [2].

In this work, the overall objective is to analyze and propose methods to deal with imbalanceness in the hierarchical classification scenarios. We are concerned in how the imbalanceness in a dataset can affect the classification results and, in addition, how to pre-process the dataset by using resampling techniques in order to minimize the imbalance issues. In order to meet the general objective of this work, we outline the following specific objectives:

1) Investigate the impacts of binary/multi-class and multi-label resampling methods on hierarchical datasets.
2) Propose metrics to measure the imbalanceness issues in the different types of hierarchical classification problems.
3) Propose techniques to deal with imbalanceness in hierarchical classification problems considering the different classification approaches.
4) Propose and investigate the use of the novel resampling measures and approaches in a real world hierarchical classification case study.

II. The use of flat resampling in imbalanced hierarchical datasets

This was the starting point, or baseline, in order to understand the existing resampling algorithms and how they could be used to deal with the imbalance issues in hierarchical classification datasets. It is important to observe that there are no concerns regarding the depth of the prediction at this point, that is, the resampling algorithms do not distinguish hierarchical classification problems with partial or full depth of prediction, since they deal with the label paths ignoring their hierarchical structure. It means that the resampling methods will totally ignore that two labels can be somehow related (Ex.: A/B and A/B/C).

A. Discussions

With these investigations we were able to answer following research questions:

- Can we apply the flat resampling algorithms in hierarchical datasets? *Yes.*
- Can the binary resampling algorithms improve the classification results in the hierarchical datasets with single paths? *Partially.*
- Can the multi-label resampling algorithms improve the classification results in the hierarchical datasets with multiple paths? *No.*

III. A label path conversion strategy

This was the first proposal towards the design of specific schemas to deal with the imbalance issue in hierarchical datasets. The main contributions are two conversion algorithms (HMC → ML and ML → HMC). The algorithms HMC ↔ ML are able to convert the label paths into multi-label formats, so we can apply multi-label resampling algorithms in the hierarchical datasets.

[1] This work relates to a Ph.D. thesis.

A. Hierarchical to Multi-Label Conversion

The main idea here is to group (for each instance) its labels paths into an unique labelset.

B. Multi-Label to Hierarchical Conversion

The first important observation regarding this conversion is that the algorithm needs the label hierarchy as input. The main idea is: For each label in the instance labelset, the algorithm will "walk through" the labels hierarchy, identifying the longest label path ending with the given label.

C. Discussions

Considering this topic, we could investigate the following research questions:

- Can we develop a method to convert a hierarchical dataset into a multi-label dataset without losing the labels relationships information? *Yes*.
- Can we measure the imbalanceness of a hierarchical dataset in a global way? *Yes*.
- Can the label path conversion strategy increase the classification results? *Yes*.

IV. Measuring imbalance in Hierarchical Datasets

The measurement of imbalanceness in datasets, known as Imbalance Ratio (IR), is usually obtained by computing a ratio between the number of samples in the majority classes and the ones associated to the minority classes. A high IR leads to a highly imbalanced dataset [1].

However, we have different classification approaches to tackle a hierarchical problem: Global algorithms and Local algorithms. As these approaches differs in how they deal with the data and how their imbalance influence in the model training process, we have proposed different metrics to measure the imbalance of the hierarchical classification problems according to the classification approach, that is, locally or globally.

A. Global measure

In Formula 1 we define $IRLP(p)$, which represents the imbalance level of a certain Label Path p. In this context, P is the set of all possible Label Paths that has at least one occurrence in any samples, P_i is the *i-th* label path, and the dataset is represented as D.

$$IRLP(p) = \frac{\max\limits_{p' \in P}(\sum_{i=1}^{|D|} h(p', P_i))}{\sum_{i=1}^{|D|} h(p, P_i)} \tag{1}$$

$$h(p, P_i) = \begin{cases} 1, & p \in P_i \\ 0, & p \notin P_i \end{cases}$$

In Formula 1, the value is 1 for the most frequent Label Path, and a greater value for the others. The higher $IRLP$ is, the larger will be the imbalance level for the Label Path.

Formula 2 defines the Mean Imbalanceness of a Hierarchical Dataset (*HMeanIR*) based on the average between the imbalanceness per label path previously presented.

$$HMeanIR = \frac{1}{|P|} \sum_{p=P_1}^{P_{|P|}} IRLP(p) \tag{2}$$

B. Local measures

In the following subsections we present novel metrics that can be used to measure imbalanceness in hierarchical datasets taking into account the local imbalance information. The idea behind these measures is to create a mechanism that can summarize and quantify the imbalanceness in the subsets created in the training step of each local classifier, considering the different local approaches (Local Classifiers per Node - LCN, Local Classifiers per Parent Node - LCPN or Local Classifiers per Level - LCL) and policies used to select the samples in the training step. Thus, we have defined three different Imbalance Ratio equations: IR_{LCN}; IR_{LCPN}; and IR_{LCL}.

For all equations, let us consider D as the hierarchical classification dataset, p as the policy chosen to select the positive/negatives samples in order to build the local classification model, n as a node/label from the hierarchy, $|L|$ as the total number of nodes/labels from the hierarchy, S_j as the j^{th} instance of the dataset, n_i as the i^{th} node from the labels hierarchy, C_n as the set of immediate children of node n, C_{n_i} as the i^{th} immediate child of n, LV as the set of levels in the label hierarchy, and N_{lv} as the set of nodes of the level lv. Moreover, for all metrics the h formulas are used in order to identify if a certain sample S_j is labeled with the given label x when using the given local approach with that specific policy p.

1) Imbalance Metrics for the LCN Approach: The LCN Imbalance Ratio for the node n with policy p, named IR_{LCN}, is defined as:

$$IR_{LCN}(n, p) = \frac{\max\limits_{x \in 0,1}(\sum_{j=1}^{|D|} h(S_j, x, p))}{\min\limits_{x \in 0,1}(\sum_{j=1}^{|D|} h(S_j, x, p))} \tag{3}$$

where:

$$h(S_j, x, p) = \begin{cases} 1 & \text{if } S_j \text{ is labeled with } x \text{ using } p, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Furthermore, we may define the Mean Imbalance Ratio ($MeanIR_{LCN}$) when using the local classifiers per node approach with the policy p as the average between the $IR_{LCN}(n, p)$ for all label nodes:

$$MeanIR_{LCN}(p) = \frac{\sum_{i=1}^{|L|} IR_{LCN}(n_i, p)}{|L|} \tag{5}$$

2) Imbalance Metrics for the LCPN Approach: The LCPN Imbalance Ratio for the node n with policy p, named $IR_{LCPN}(n, p)$, is defined as:

$$IR_{LCPN}(n, p) = \frac{1}{|C_n|^2} \sum_{i=1}^{|C_n|} \frac{\sum_{j=1}^{|D|} h(S_j, C_n, p)}{\sum_{j=1}^{|D|} h(S_j, C_{n_i}, p)} \tag{6}$$

where:

$$h(S_j, x, p) = \begin{cases} 1 & \text{if p = sib. and } S_j \text{ is labeled with } x, \\ 1 & \text{if p = exc. sib. and } S_j \text{ is labeled with } x, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Moreover, we may define the mean Imbalance Ratio when using the local classifiers per parent node approach with the policy p, named $MeanIR_{LCPN}$, as:

$$MeanIR_{LCPN}(p) = \frac{\sum_{i=1}^{|PN|} IR_{LCPN}(n_i, p)}{|PN|} \tag{8}$$

3) Imbalance Metrics for the LCL Approach: The LCL Imbalance Ratio for the level lv, named IR_{LCL}, is defined as:

$$IR_{LCL}(lv) = \frac{1}{|N_{lv}|^2} \sum_{i=1}^{|N_{lv}|} \frac{\sum_{j=1}^{|D|} h(S_j, N_{lv})}{\sum_{j=1}^{|D|} h(S_j, n_i)} \tag{9}$$

where:

$$h(S_j, x) = \begin{cases} 1 & \text{if } S_j \text{ is labeled with label } x \text{ (or in } x\text{)}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Therewithal, we may define the mean Imbalance Ratio when using the local classifiers per level approach with the policy p, named $MeanIR_{LCL}$, as:

$$MeanIR_{LCL} = \frac{\sum_{lv=1}^{|LV|} IR_{LCL}(lv)}{|LV|} \tag{11}$$

C. Discussions

Considering this topic, we could investigate the following research questions:

- Can we measure the imbalanceness of a hierarchical dataset in a global way? *Yes*.
- Can we measure the imbalanceness in the hierarchical datasets considering the LCN, LCPN and LCL approaches? *Yes*.

V. Local resampling approaches

Among the techniques to deal with hierarchical classification, the local approaches are well-known approaches in the literature. We have proposed three different resampling schemas, considering the three different local approaches (LCN, LCPN and LCL), in the following subsections.

A. Resampling Using the LCN Approach

The main idea here is to resample each binarized dataset before building the classification model for each node. The classification schema is composed of three main steps: (1) Building one binary classifier per label node in the hierarchy; (2) Classifying the test dataset; (3) Measuring the classification results with a hierarchical measure. Even though these steps are already commonly used in order to classify a hierarchical dataset with the LCN approach, the first step is further subdivided into three substeps: (1.1) Applying a previously defined policy to choose the positive/negative samples when building the classification model for a given node n; (1.2) Applying a flat binary classification resampling algorithm in the binarized training dataset; (1.3) Using a flat single-label classification algorithm to build the classification model for node n. It is important to observe that the proposed approach is specifically embedded into the step 1.2, in which a binary resampling process is applied into the training dataset.

During the testing phase (step 2), we use a top-down approach to predict the hierarchy of labels for a new sample, avoiding inconsistencies in class prediction at the different levels. It means that given an unknown sample, the idea is to walk down into the model tree predicting if the sample belongs to each label from the hierarchy. This way, we have to use a threshold to define if we must consider a sample belonging to a certain label or not. It is important to note that we only keep moving down the next node of the model tree if the sample is labeled with the previous node.

B. Resampling Using the LCPN Approach

Similarly to LCN, the classification schema is also composed of three main steps: (1) Building a classification model for each parent node in the labels hierarchy; (2) Classifying the test dataset using a top-down approach; (3) Measuring the results with a hierarchical measure. Such as in LCN, there are three substeps in the model building phase, where in substep 1.1 a policy has to be chosen (in this scenario only siblings or exclusive siblings are allowed). The proposed resampling phase is also embedded into substep 1.2. It might be observed that, as the LCPN approach creates multi-class problems for each parent node and the classic resampling approaches are used to work in binary class problems, we have to apply an O-A-A or an O-A-O approach. These techniques decompose a multi-class classification problem into a series of binary subproblems, so we can apply a binary resampling algorithm in each one of them. Finally, on substep 1.3, the parent node model is built considering a single-label classification algorithm.

In substep 1.1, differently to the LCN approach, we may use only two different policies to choose the samples from the train dataset in order to build the classification model for a certain parent node n.

C. Resampling Using the LCL Approach

Similarly to the previously schemas, the classification is composed of three main steps: (1) Building a classification model for each level in the labels hierarchy; (2) Classifying the test dataset; (3) Measuring the results with a hierarchical measure. The first important difference from the other approaches is that even using a top-down technique, the classification may predict labels with an inconsistency between the classes from different levels, which has to be removed later.

Step 1 is also subdivided into three substeps and, such as in LCN and LCPN approaches, the proposed resampling schema is embedded into substep 1.2. In this substep we have also an O-A-A or an O-A-O approach to decompose the multi-class classification problems per level into a series of binary sub-problems and then apply the classic binary resampling algorithms.

On the contrary of the LCN and LCPN approaches, we do not have different policies to apply on substep 1.1, since we must select all samples labelled with the labels from the level that we are building the classifier to.

D. Discussions

Considering this topic, we could investigate the following research questions:

- Can the flat resampling algorithms improve the results in the LCN, LCPN and LCL approaches? *Yes*.
- Can the proposed local resampling schemas reduce the imbalanceness considering the proposed metrics, that is, IRLCN, IRLCPN and IRLCL? *Yes*.
- Does the policy used to select the subset samples during the local training steps of LCN and LCPN influence in the resampling? *Yes*.

VI. Global resampling approaches

We have proposed three novel resampling algorithms for hierarchical classification problems: (i) Hierarchical Random Oversampling (HROS); (ii) Hierarchical Random Undersampling (HRUS); and (iii) Hierarchical Synthetic Oversampling Technique (HSMOTE).

In order to design these novel resampling algorithms, we have considered two different variants of hierarchical problems described in [3]: number of paths (defined in the literature as Ψ) and depth of the paths (defined in the literature as Φ).

A. Finding the majority and minority label paths

Before proposing any resampling algorithm for hierarchical datasets, we have to establish a mechanism to identify the majority and minority classes. In hierarchical problems, the classes are represented by label paths in the tree taxonomy instead of individual labels. Our idea here is to find the majority and minority labels paths in the dataset based on their imbalance ratio, which are calculated with Formulas 1 and 2. Thus, the majority label paths are those whose IRLP is lower then HMeanIR, while the minority label paths are those whose IRLP are greatter then HMeanIR.

B. Resampling partial depth hierarchical classification problems

The full depth hierarchical classification problems can be directly handle by the resampling algorithms. However, in the partial depth problems, the instances may be associated with one or more label paths with full or partial depth in the label tree. The challenge of resampling this kind of data is that when creating or removing samples from children nodes, the number of samples from parent label nodes will be indirectly increased or removed. This problem does not affect full depth hierarchical classification problems because there are no samples labeled exclusively with internal nodes.

In order to deal with this issue, we proposed a technique to process the instances in a "bottom-up order", recalculating the majority/minority paths after each loop of the resampling process. Figure 1 shows an example of the proposed method for the HROS algorithm. For this example the resample process takes 3 steps (starting on the leaf nodes until reach the root). The example dataset is composed of 85 samples and a label tree with 9 nodes. We simulated the application of an oversampling method with an increase rate of 15%.

C. Hierarchical Random Oversampling/Undersampling (HROS/HRUS)

Considering the previously described strategies, we may propose the HROS and HRUS by finding the minority/majority label paths and randomly duplicating/removing their samples in order to achieve a label path distribution corresponding to the resize rate chosen by the user.

D. Hierarchical Synthetic Oversampling Technique (HSMOTE)

When proposing a synthetic oversampling algorithm, there are five main aspects to solve:

1) Minority instances selection: A criterion to define and select which label paths belong to the minority set of paths has to be established. Here, we can use the criteria described in subsection *A*.
2) Different kinds of hierarchical problems and relationship between the labels: The resampling process has to be investigated in each hierarchical classification scenario (full or partial depth prediction and single or multiple paths) and a mechanism to deal with the labels hierarchy (mainly in partial depth problems) has to be defined. Here, we can use the criteria proposed subsection *B*.
3) Nearest neighbor search: Given an instance that belongs to a minority label path, the algorithm has to search its nearest neighbors which will be used to generate the synthetic sample. Here, we can use the same strategy used in MLSMOTE algorithm [4].
4) Feature set generation: After selecting the neighbors, the set of features for the synthetic sample is obtained through interpolation techniques. Here, we can also use the same strategy used in MLSMOTE algorithm [4].
5) Synthetic labelset production: Since we have different kinds of hierarchical classification problems, the production of synthetic path(s) also depends on the type of the problem.

Regarding the association of label(s) to the synthetic instance, in the classic SMOTE [5], as it deals with binary/multi-class problems, the label of the sample selected from the minority set is cloned to the synthetic sample. In MLSMOTE, as it handle multi-label problems, [4] proposed the use of three label combinations techniques with the neighbors labels to solve the issue producing a new labelset: Intersection, Union and Ranking. As HSMOTE deals with hierarchical classification problems, which can has single or multiple paths (defined as Ψ) and be either full or partial depth (defined as

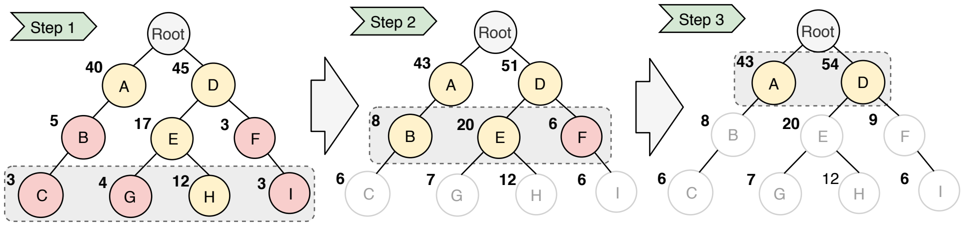

Fig. 1. An example of application of the HROS in a partial depth problem with 85 instances. The nodes marked with a circled dashed are being processed at the certain step and the red nodes represent the label paths belonging to the minority set.

Φ), we designed the following possibilities for the label(s) generation, according to the problem's taxonomy:

- Single Path Problems:
  - Full Depth (FD):
    - Clone: The label path of the seed sample, that is, the instance selected from the set of minority paths, is cloned to the synthetic sample.
  - Partial Depth (PD):
    - Clone: The same as in FD, that is, the label path of the seed sample is cloned.
    - Longest Common Path: The longest common path among the neighbors is chosen as the label path for the synthetic sample.
- Multiple Path Problems (FD or PD):
  - Union: All label paths that appear in the reference instance or any of its neighbors are used as the synthetic labelset.
  - Intersection: The label paths that appear in the reference instance and the neighbors are used as the synthetic labelset.
  - Ranking: We count the number of occurrences of each label path in the reference sample and its neighbors and those which are present in half or more of the instances are considered as labelset of the synthetic sample.

However, when dealing with partial depth problems, we have to handle the following issues in each label combination criteria:

- Union: The combination of a partial depth path and its full depth path results in the full depth path, since the partial depth path belongs to the full path.
- Intersection: The combination of two label paths can lead to a common partial depth path between them.
- Ranking: All label paths present in the samples (partial and full depth) are ranked according to their frequency and, during this ranking step we have to take into account when a partial depth path is present in a full depth path.

E. *Discussions*

When designing these methods, we were able to answer the following research questions:

- Can we define a way to retrieve the majority and minority sets of label paths in a hierarchical dataset? *Yes*.
- Can we deal with the different types of hierarchical problems? *Yes*.
- Can we produce synthetic sets of label paths by combining neighbors' instances? *Yes*.

VII. The COVID-19 identification in CXR images case of study

As our main case of study, we aimed to explore the identification of different types of pneumonia caused by multiple pathogens using CXR images with textural features. Specifically, we considered pneumonia caused by viruses (COVID-19, SARS, MERS and Varicella), bacteria (Streptococcus) and fungus (Pneumocystis). These pathogens (labels) were hierarchically organized according to their biological relationships.

A. *General overview*

To better understand the case of study, Figure 2 shows a general overview of the classification schema, considering: The feature extraction process (Phase 1), the Early Fusion technique (Phase 2), the data resampling (Phase 3), the classification (Phase 4) and Late Fusion technique (Phase 5). It should be noted the reasoning behind this naming schema is as follows: Phases 1 and 2 are the same though all configurations, while Phases 3, 4 and 5 may change according to the resampling approach (that is why they are presented in dashed lines in the Figure 2). It should be noted that all the techniques proposed in this Thesis were tested on Phase 3.

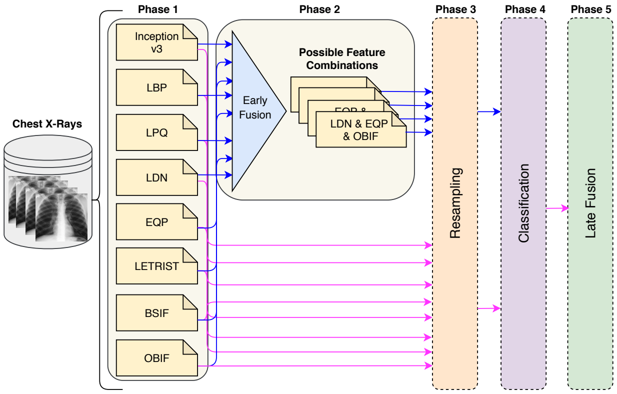

Fig. 2. A general classification schema for the COVID-19 identification in CXR images. While the blue lines represents the early fusion connections, the pink lines are used for the late fusion and results without fusion.

B. *The dataset*

The pneumonia CXR dataset used in the experiments (named RYDLS-20) is another proposal of this section. Table I shows the data distribution of this dataset.

TABLE I
RYDLS-20 SAMPLES DISTRIBUTION.

Label Path	*#Samples*	*#Train*	*#Test*
Normal	1,000	700	300
Pneumonia/Acellular/Viral/Coronavirus/COVID-19	90	63	27
Pneumonia/Acellular/Viral/Coronavirus/MERS	10	7	3
Pneumonia/Acellular/Viral/Coronavirus/SARS	11	8	3
Pneumonia/Acellular/Viral/Varicella	10	7	3
Pneumonia/Celullar/Bacterial/Streptococcus	12	9	3
Pneumonia/Celullar/Fungus/Pneumocystis	11	8	3

C. Discussion

When investigating this topic, we were able to answer the following research questions:

- Can we use textural features to recognize the different pathogens CXR images? *Yes*.
- Can the hierarchical classification schema perform better than a classic flat classification schema for the COVID-19 identification task? *Yes*.
- Can the proposed resampling schemas improve the baseline classification results? *Yes*.
- Can the pattern recognition techniques differentiate the types of pathogens causing pneumonia? *Partially*

VIII. Concluding Remarks and Contributions

In this work, we proposed novel approaches to deal with the imbalanceness issue for different types of hierarchical classification problems. In the following, we describe the scientific and technical contributions achieved with this work. Besides, we also present the social impact of this work.

A. Scientific Contributions

In Table II we present a summary of the papers that were sent for publication during the development of this work. We present the paper's reference (or date of submission), the relation of the paper and the Section(s) of this document that describe its contributions, the Impact (Impact Factor (IF) for journals and h5-index (H5i) for conferences), the status of the publication and the current number of citations.

TABLE II
PAPERS DEVELOPED DURING THIS RESEARCH.

Ref.	*Sections*	*Venue*	*Impact*	*Status*	*Citations**
[6]	2	ICME	H5i: 30	Published	5
[7]		FLAIRS	H5i: 16		5
[8]		IJCNN	H5i: 46		3
[9]		Multimedia Tools and Application	IF: 2.31		2
[10]		Neurocomputing	IF: 4.44		10
[11]	3 and 4	ICTAI	H5i: 19		5
[12]	2, 3, 4, 5 and 6	Computer Methods and Programs in Biomedicine	IF: 3.63		185
[13]	4 and 5	Data Mining and Knowledge Discovery	IF: 2.63		2
[14]	6	Information Sciences	IF: 5.91		0
-		Journal of Machine Learning Research	IF: 5.92	Under Review	-

* Citations obtained from Google Scholar on August 30, 2021.

B. Technical Contributions

During the development of this work, two frameworks were designed: (i) The Imb-Mulan , a multi-label imbalance learning library; and (ii) The hierarchical-imblearn , a hierarchical imbalanced learning library.

Beyond the frameworks, we proposed many novel datasets in order to investigate the effects of the proposed classification and resampling approaches. In Table III, we present a brief summary of the proposed datasets and where to find them.

TABLE III
NOVEL DATASETS PROPOSED IN THIS THESIS.

Dataset	*Type of Classification*	*Domain*	*Link for Download*
RYDLS-20	Single-Label and Hierarchical	Medical	https://bit.ly/rydls-20
P-TMDB	Multi-Label	Movie	https://bit.ly/p-tmdb
FMA90k		Music	https://bit.ly/fma-90k
FMA-SL	Single-Label		https://bit.ly/fma-sl
BRMD			https://bit.ly/brmd
Hier-CAL500	Hierarchical		https://bit.ly/h-imb-db
Hier-Emotions			
Hier-FMA-MFCC			
Hier-FMA-SL-LBP			
Hier-FMA-SL-SSD			
Hier-Enron		Text	
Hier-Birds		Animal	

C. Social and Media Impact

This work was partially developed during the breakthrough of the COVID-19 pandemic. Considering this context, we developed a method to identify COVID-19 and other pneumonia pathogens in CXR images. Our work was one of the first studies published in the literature addressing this issue [12]. Given the importance of the topic and the timely publication of this contribution, it has attracted the attention of researchers, society, and media vehicles.

In order to give an overview of the repercussions, in Table IV we present a summary of the main reports concerning our work in the TV, Radio and Magazines. It is worth mentioning that the report from the Agencia Brasil was republished by over a 100 online news agency websites, including Valor Econômico, Época and Istoé.

TABLE IV
SUMMARY OF THE MAIN MEDIA REPORTS CONCERNING OUR WORK.

Report Link	*Repercussion Level*	*Venue*	*Media Type*
https://bit.ly/physics-covid19	International	Physics	Magazine
https://bit.ly/gazeta-covid19	National	Gazeta do Povo	Magazine
https://bit.ly/capes-covid19	National	CAPES	Youtube Channel
https://bit.ly/cnnbrasil-covid19	National	CNN Brasil	TV
https://bit.ly/jovempan-covid19	National	JovemPan News	Radio
https://bit.ly/3r12JLq	National	SuperAcesso	Magazine
https://bit.ly/agenciabrasil-covid19	National	Agencia Brasil*	Magazine
https://bit.ly/cbn-covid19	State	CBN Curitiba	Radio
https://bit.ly/lightnews-covid19	State	Transamérica	Radio
https://bit.ly/rpc-covid19	State	RPC Parana	TV
https://bit.ly/cbn-mga-covid19	Local	CBN Maringá	Radio
https://bit.ly/ric-covid19	Local	RIC Maringá	TV
https://bit.ly/rpc-mga-covid19	Local	RPC Maringá	TV

* This report was republished by over a 100 online news agency websites.

References

[1] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[2] C. Xu and X. Geng, "Hierarchical classification based on label distribution learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu, USA: AAAI, 2019, pp. 5533–5540.

[3] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[4] F. Charte, A. Rivera, M. del Jesus, and F. Herrera, "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation," *Knowledge-Based Systems*, vol. 89, pp. 385–397, 2015.

[5] N. Chawla, K. Bowyer, L. Hall, and P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] R. M. Pereira and C. N. Silla Jr, "Using simplified chords sequences to classify songs genres," in *Proceedings of IEEE International Conference on Multimedia and Expo*. Hong Kong: IEEE, 2017, pp. 1446–1451.

[7] V. D. Valerio, R. M. Pereira, Y. M. G. Costa, D. Bertoini, and C. N. Silla Jr, "A resampling approach for imbalanceness on music genre classification using spectrograms," in *Proceedings of the International Florida Artificial Intelligence Conference*. Melbourne, USA: AAAI, 2018.

[8] R. M. Pereira, Y. M. G. Costa, R. L. Aguiar, A. S. Britto Jr, L. E. S. Oliveira, and C. N. Silla Jr, "Representation learning vs. handcrafted features for music genre classification," in *Proceedings of the International Joint Conference on Neural Networks*. Budapest, Hungary: IEEE, 2019, pp. 1–8.

[9] R. B. Mangolin, R. M. Pereira, A. S. Britto Jr, C. N. Silla Jr, V. D. Feltrim, D. B. Goncalves, and Y. M. G. Costa, "A multimodal approach for multi-label movie genre classification," *Multimedia Tools and Applications*, vol. 79, no. 43, pp. 1–30, 2020.

[10] R. M. Pereira, Y. M. G. Costa, and C. N. Silla Jr, "MLTL: A multi-label approach for the tomek link undersampling algorithm," *Neurocomputing*, vol. 383, pp. 95–105, 2020.

[11] ——, "Dealing with imbalanceness in hierarchical multi-label datasets using multi-label resampling techniques," in *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*. Volos, Greece: IEEE, 2018, pp. 818–824.

[12] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla Jr, and Y. M. G. Costa, "COVID-19 identification in chest x-ray images on flat and hierarchical classification scenarios," *Computer Methods and Programs in Biomedicine*, vol. 194, pp. 1–28, 2020.

[13] R. M. Pereira, Y. M. Costa, and C. N. Silla, "Handling imbalance in hierarchical classification problems using local classifiers approaches," *Data Mining and Knowledge Discovery*, vol. 35, pp. 1564–1621, 2021.

[14] ——, "Toward hierarchical classification of imbalanced data using random resampling algorithms," *Information Sciences*, vol. 578, pp. 344–363, 2021.