A co-training -based approach for the hierarchical multi-label classification of research papers

Abir Masmoudi, Hatem Bellaaj, Khalil Drira, Mohamed Jmaiel

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A Co-Training-based Approach for the Hierarchical Multi-label Classification of Research Papers

Abir Masmoudi*[1] | Hatem Bellaaj[1] | Khalil Drira[2] | Mohamed Jmaiel[1]

[1]ReDCAD laboratory , University of Sfax, Sfax, Tunisia
[2]LAAS-CNRS, Univ. de Toulouse, Toulouse, France

Correspondence
*Abir Masmoudi, Email: abir.masmoudi@redcad.org

Summary

This paper focuses on the problem of the hierarchical multi-label classification of research papers, which is the task of assigning the set of relevant labels for a paper from a hierarchy, using reduced amounts of labeled training data. Specifically, we study leveraging unlabeled data, which are usually plentiful and easy to collect, in addition to the few available labeled ones in a semi-supervised learning framework for achieving better performance results. Thus, in this paper, we propose a semi-supervised approach for the hierarchical multi-label classification task of research papers based on the well-known Co-training algorithm, which exploit content and bibliographic coupling information as two distinct papers' views. In our approach, two hierarchical multi-label classifiers, are learnt on different views of the labeled data, and iteratively select their most confident unlabeled samples, which are further added to the labeled set. The success of our suggested Co-training-based approach lies in two main components. The first is the use of two suggested selection criteria (i.e. Maximum Agreement and Labels Cardinality Consistency) that enforce selecting confident unlabeled samples. The second is the appliance of an oversampling method that rebalances the labels distribution of the initial labeled set, which reduces the reinforcement of the label imbalance issue during the Co-training learning. The proposed approach is evaluated using a collection of scientific papers extracted from the ACM digital library. Performed experiments show the effectiveness of our approach with regards to several baseline methods.

KEYWORDS:
Hierarchical Multi-label classification,Co-training, Semi-supervised learning, Imbalanced data, Research papers classification

1 | INTRODUCTION

The classification of research papers becomes a necessity for ensuring an easy search, access and retrieval by scientists. In many real world applications, papers are assigned to many labels simultaneously that are arranged in a hierarchical structure. As an example, the ACM classification tree [1], which defines the concepts of the computer science domain, is a typical case of hierarchical schemas used for the classification of scientific publications. This type of classification tasks is referred to as the Hierarchical Multi-label Classification (HMC). In HMC, the predicted relevant labels for a paper could belong to several paths in the hierarchy. Furthermore, the hierarchy constraint should be satisfied: i.e. when a label is assigned, all its parent labels are equally assigned. Hence, HMC methods, which take into account the hierarchy structure, are necessary for ensuring the consistency of the assigned labels with the hierarchy constraint. Proposed HMC methods in the literature are divided into two major types: global methods and local methods. The former train a single global classifier that discriminates the relevant labels for a given instance from the whole

[1]https://dl.acm.org/ccs/ccs_flat.cfm

label hierarchy (Cerri, Barros, & de Carvalho 2012). The latter build a separate local classifier for each label, parent label or hierarchy level. Then, the local classifiers' outputs are combined to deduce the final assigned labels while considering the hierarchy structure, often based on a top-down or a bottom-up induction mechanism.

These HMC methods require using large amounts of labeled training data to achieve good classification performances. However, because of the difficulty of the manual labeling task, there are often few labeled data. Semi-Supervised Learning (SSL) (Chapelle & Zien 2006) is an effective strategy to address this problem by exploiting unlabeled data, which are usually plentiful and easy to collect, in addition to the labeled ones, for improving the classification performances. In particular, Co-training (Blum & Mitchell 1998), which is originally proposed by Blum and Mitchell, is one of the most popular SSL algorithms. It assumes the presence of two different data' views, which are described by disjoint feature subsets, often represented by two feature vectors. Ideally, each view is sufficient for the learning step, and the views are conditionally independent from each other given the class labels. Firstly, two classifiers are separately trained on the initial labeled data using different feature sets corresponding to distinct views. Then, each classifier selects its most confidently predicted unlabeled samples, which are further added (with their assigned labels) to the labeled set, and removed from the unlabeled set. The classifiers are retrained on the updated labeled set, and the process iterates until the maximum number of iterations is reached.

Blum and Mitchell have successfully applied the Co-training algorithm for the binary classification task, wherein samples are associated with a single label. Then, many studies have applied the Co-training algorithm in different domains. However, they mostly consider either the single-label classification (M.-L. Zhang & Zhou 2011), or the flat multi-label classification (Zhan & Zhang 2017) tasks, and thus ignore the structural relationships between labels. Hence, these studies cannot effectively deal with the difficulties of applying the Co-training algorithm in HMC settings. Specifically, in HMC settings, there are many issues that hamper the well-functioning of the Co-training algorithm. First, classifiers, which are learnt using small labeled training datasets, are not enough accurate to make strong hypotheses on the different labels. Therefore, it is expected that a large proportion of unlabeled samples to be misclassified, e.g. assigned with one or many irrelevant labels. Second, error propagation is a common problem in HMC settings since misclassifications can spread from higher to lower levels of the hierarchy (Irsan & Khodra 2019), or inversely. Hence, identifying which samples are associated with confident labels is a more difficult task in HMC settings. Third, in typical HMC scenarios, the number of training samples for labels at lower hierarchical levels is often smaller than for labels at the higher hierarchical levels (Otero, Freitas, & Johnson 2010). Moreover, labels at the lower levels are usually characterized by a strongly skewed distribution (Chen & Hu 2012). In such imbalanced learning scenarios, classifiers are biased towards the majority labels while ignoring the minority ones. As a result, the label imbalance issue may be exaggerated with the progress of the Co-training learning, which degrades the performance scores (Xing, Yu, Domeniconi, Wang, & Zhang 2018). All these factors render challenging the success of the Co-training algorithm in HMC settings.

In this paper, we propose a Co-training-based approach that deals with the aforementioned issues to achieve better results by the use of unlabeled data in HMC settings. Our proposed approach, which is applied on scientific papers classification task, exploits content and bibliographic coupling information as two distinct papers' views. It comprises three main steps: (1) rebalancing, (2) learning, and (3) classification (as shown in Figure 1). In the first step, the initial labeled set is rebalanced using a Multilabel Synthetic Minority Oversampling Technique (MLSMOTE) (Charte, Rivera, del Jesus, & Herrera 2015) that generates synthetic samples, associated with the minority labels. Then, in the learning step, the classifiers are firstly trained on the rebalanced labeled dataset while exploiting separately different papers' views. Afterwards, the classifiers predict the labels of unlabeled samples to select a set of their most confidently predicted ones. Those latter are discovered using a novel selection mechanism that relies on two suggested criteria, which are: Maximum Agreement and Labels Cardinality Consistency (detailed in section 3.1.3). The classifiers are then retrained on the updated labeled set with the confident predictions of each classifier, and the process iterates until the stopping condition is satisfied (i.e. the maximum number of iterations is reached). Finally, at the classification step, the final classifiers outputs are combined via the consensus principle to infer the final assigned labels for unseen test examples.

Performed experiments have validated the effectiveness of the suggested approach in improving the performance scores by the use of unlabeled data, and prove the necessity of combining the two suggested criteria for achieving a successful Co-training learning. Experiments have also shown the advantage of the incorporation of MLSMOTE in reducing the reinforcement of the label imbalance issue with the progress of the Co-training learning.

The remainder of this paper is organized as follows. Section 2 reviews related work in the literature on scientific papers classification. Section 3 provides an overview of the proposed approach, and describes its different components. Section 4 presents an evaluation study of our approach along with a discussion about the obtained results. Section 5 concludes with some future research directions.

2 | RELATED WORK

In this section, we firstly give an overview on existing techniques for solving the HMC task. Then, we present state-of the art approaches on scientific papers classification.

2.1 | Hierarchical Multi-label Classification

Many studies have been proposed in the literature for the HMC task. Initially, flat classification methods are used for solving the HMC task, typically assigning labels that are leaf nodes in the hierarchy (Costa, Lorena, Carvalho, Freitas, & Holden 2007). The main disadvantage of these approaches is that they do not explore information about parent-child relationships, which limits their efficiency.

Several sophisticated HMC methods are further proposed to consider the hierarchy structure information, which are divided into local and global approaches. According to (Silla & Freitas 2011), local approaches rely on three main strategies for exploring the local information present in the hierarchy during the training phase: a Local Classifier per Node (LCN), a Local Classifier per Parent Node (LCPN) and a Local Classifier per Level (LCL) strategy. In LCN strategy, a local binary classifier is trained for each node (i.e. label) in the hierarchy. In LCPN, a multi-label classifier is learnt for each parent label to distinguish between its children labels. In LCL, a multi-label classifier is induced for each hierarchical level that predicts labels present in its associated level. As an example, (Cerri, Barros, & De Carvalho 2014) propose an LCL-based local approach that trains a Multi-Layer Perceptron (MLP) neural network for each hierarchical level. Predictions made for training samples in a specific level are taken as additional features for the MLP network learner at the lower level. Then, at the test phase, a top-down strategy that surfs the label hierarchy from root to leaf nodes is adopted for deducing the final assigned labels. (Feng, Fu, & Zheng 2018) propose an LCN-based HMC method that firstly applies a negative instances selecting policy and an oversampling technique for reducing the label imbalance problem in initial training datasets. Then, a particular MLP classifier is learnt for each label in the hierarchy. A post-processing method based on the Bayesian network is used to ensure the consistency of the final predicted labels with respect to the hierarchy constraint. Several kernel-based studies (e.g.(Rousu, Saunders, Szedmak, & Shawe-Taylor 2006; Valentini 2010)) adopting an LCN strategy were proposed. For instance, (Valentini 2010) suggests using a set of probabilistic SVMs, which are learnt for each label separately. At the testing phase, final labels are obtained using a bottom-up strategy that recursively propagates positive predictions towards the higher level labels (i.e. their parents, and all their ancestors), which enforces labels consistency. In (A. Santos & Canuto 2014), the authors suggest a local method, called HMC-RAKEL, which adapts the multi-label method, RAKEL (Tsoumakas & Vlahavas 2007), to take into account the hierarchical dependencies between labels. Specifically, it adopts an LCPN strategy at the training step while using RAKEL as a basic multi-label learner. Then, labels' predictions of unseen test examples are induced by the classifiers in a top-down manner.

In (Tsoumakas, Katakis, & Vlahavas 2008), the authors suggest HOMER, a meta-algorithm for multi-label learning, that firstly builds a hierarchy of label sets from a set of disjoint labels using a balanced clustering algorithm. Using the constructed label hierarchy, it trains a multi-label classifier for each parent label having more than one child label, and applies a top-down induction mechanism at the test phase. As revealed in (Tsoumakas et al. 2008), HOMER can be extended to work with an existing label hierarchy, which makes it a local HMC method.

As for global approaches, they build a single classifier that copes with all labels in the hierarchy. One of the first methods, which is called HMC4.5 (Clare & King 2003), makes use of decision trees for constructing the global model. More precisely, HMC4.5 adapts the well-known decision tree algorithm C4.5 (Quinlan 2014) to the HMC settings based on a modified computation of the class entropy. The novel computation method considers the individual entropy of each class separately, and also the hierarchical relationships between labels.

Other studies (Blockeel, Bruynooghe, Džeroski, Ramon, & Struyf 2002; Blockeel, Schietgat, Struyf, Džeroski, & Clare 2006) propose the use of the concept of Predictive Clustering Trees (PCTs) for HMC tasks. Specifically, they suggest a global method, called Clus-HMC, that builds a single tree for all labels at once. In (Vens, Struyf, Schietgat, Džeroski, & Blockeel 2008), Clus-HMC method is compared with two local HMC methods. The first, named Clus-SC, trains a decision tree for each label independently, while the second, named Clus-HSC, builds a separate decision tree for each parent label in the hierarchy. Experiments prove the efficiency of Clus-HMC over compared methods, and show that using a classifier per each parent node leads to better performances than at each node separately. Still based on PCTs, (Valentini 2010) propose using ensembles of PCTs induced by Clus-HMC method. Specifically, their proposed method, called Clus-HMC-Ens, adopts the Bagging technique (Breiman 1996), which makes bootstrap replicates of the training data, and trains a single decision tree classifier for each bootstrap. Conducted experiments show the advantage of using ensembles on several biological datasets.

(Otero et al. 2010) propose an ant colony optimization algorithm for HMC tasks. Their proposed algorithm, called hAnt-Miner, discovers HMC rules in the form of IF antecedent THEN consequent rule by synchronizing two ant colonies that identify the rules antecedents and consequents. Experiments show that hAnt-Miner leads to competitive results to Clus-HMC method. (Cerri et al. 2012) propose a global method with a genetic algorithm for solving the HMC task, named HMC-GA, by creating HMC rules. Specifically, HMC-GA evolves the antecedents of HMC rules using a fitness function that considers the sample coverage. Then, a set of optimal rules is used for determining the consequents of the rules, i.e. the classes to which samples belong to. In (Cerri, Basgalupp, Barros, & de Carvalho 2019), HMC-GA method is enhanced by using novel fitness functions and genetic operators that create relational and propositional rules. (Baker & Korhonen 2017) propose a global HMC method that initializes the final hidden layer of a CNN neural network model in a way that label co-occurrence relations (e.g. hypernymy) are well leveraged. In their method, a binary cross-entropy loss function is used to capture co-occurrence information between labels in the training data. Experiments show the usefulness of their approach on two HMC tasks in the biomedical domain.

The main disadvantage of global methods is the increased complexity of the built global model, especially in cases of large-scale datasets (Sun, Kudo, & Kimura 2016) that leads to high computational cost and less effectiveness.

Unfortunately, the efficiency of HMC methods (i.e. local and global approaches) may severely degrade in real world situations because of the scarcity issue of labeled training data. Applying SSL algorithms that can use any HMC method as a base learner is an interesting option since it may achieve improved performances with regards to the supervised baseline by exploiting unlabeled data. However, there is a clear lack of SSL proposals in the literature targeting the aforementioned issue. Most suggested semi-supervised studies (e.g. (Metz, Freitas, et al. 2009)) consider the hierarchical single-label classification task, and do not deal the multi-label case, i.e. a sample can be associated with many labels that could belong to different paths in the hierarchy.

To our knowledge, the presented work in (A. Santos & Canuto 2014) is the only study suggesting semi-supervised methods for the HMC task. Specifically, the authors have proposed multiple variants of the semi-supervised self-training algorithm using different HMC classifiers (e.g. HMC-RAkEL). In their suggested methods, self-training is applied in a pre-processing phase wherein a single classifier is used to label a set of randomly chosen unlabeled samples in each iteration, which are further added to the labeled set. Experiments have validated the advantage of applying SSL on the performance results.

Unlike (A. Santos & Canuto 2014), in this work, we propose a variant of the multi-view semi-supervised Co-training algorithm, which is adapted to the specificity of the HMC task. Our approach benefits from the multi-view information in data to improve the performances of two hierarchical multi-label classifiers, which are learnt using separately distinct data' views. The well-behaving of the proposed Co-training-based method is due to: (i) the use of a novel mechanism for confident samples selection based two suggested criteria (i.e. Maximum Agreement and Labels Cardinality Consistency) that guarantees achieving better results by the use of unlabeled data, and (ii) the incorporation of an oversampling technique called MLSMOTE (Charte et al. 2015) that reduces the exaggeration of the label imbalance issue during the Co-training learning.

2.2 | Scientific papers classification

Several machine learning-based approaches are proposed in the literature for the automatic classification of research papers according to a set of predefined labels, which are conducted either in a supervised or semi-supervised manner.

1)Supervised approaches: They require the use of only labeled training data for classifier training. Some studies (Łukasik, Kuśmierczyk, Bolikowski, & Nguyen 2013; Roul & Sahoo 2017; A. P. Santos & Rodrigues 2009; Surkis et al. 2016; Vogrinčič & Bosnić 2011; Zhao et al. 2018) use text-based techniques for mining papers' content information, which can be extracted from their different sections (i.e. abstract, title, keywords, and full texts). Their main assumption is that papers conducted in the same research fields, use approximately the same terms/keywords for describing the papers research goals, backgrounds, and conclusions. Most of these content-based approaches rely on the Bag of Words (BoW) model to transform the original papers' contents into a suitable representation for machine-learning classifiers. The BoW model builds a feature space based a vocabulary of terms that contains distinct words from the papers collection. In BoW, each paper in the collection is represented as a vector of term weights, where each dimension corresponds to a specific term feature from the vocabulary terms, and its associated weight indicates the presence or the absence of the term feature in the paper, which can be quantified by its frequency of occurrence. Thus, the similarities between papers can be easily computed based on their BoW representations by considering their overlapping term features. In (Zhao et al. 2018), the authors have also investigated the use of Glove word embedding models (Pennington, Socher, & Manning 2014) for generating the feature representations of papers in the biomimicry domain based on their abstract words.

Other studies (Couto et al. 2006 2010) perform an analysis of citation links between papers to infer the similarity degree between them, which facilitates the discovery of paper' topics. Citation links have two major kinds: in-links and out-links. For a given paper, in-links refer to its citing papers while out-links points to its cited references. The analysis of citation links can be performed using bibliometric measures (e.g. Co-citations (Small 1973), Bibliographic Coupling (BC) (Kessler 1963)) that search indirect citation links. The idea behind these measures is that when two papers are co-cited by many other papers or citing a large number of common papers, they may be related to the same topic. For example, the Co-citations measure (Small 1973) defines the proximity between two papers as the number of their common in-link citations (i.e. shared citing papers). As for the BC measure (Kessler 1963), it estimates the similarity between two papers based on the quantity of their common out-link citations (i.e. common cited references). The pairwise papers similarities provided by a specific bibliometric measure are used to obtain a symmetric similarity matrix wherein each cell contains the bibliometric similarity score between a pair of papers from a collection, and each row corresponds to the citation-based feature vector of a paper. The similarity matrices are taken as inputs to the machine learning classifiers. In (Couto et al. 2006 2010), comparative studies are made to assess the performance of classifiers that use different bibliometric similarity measures (i.e. Co-citations (Small 1973), Bibliographic Coupling (BC) (Kessler 1963), Amsler (Amsler 1972)) on papers classification task. In these studies, classifiers using BC measure yield to better classification results than classifiers using the Co-citations measure. Moreover, classifiers using Amsler measure, which is a linear combination of Co-citations and BC measures, achieve slightly improved results than the ones using BC measure. The obtained results are justified by the fact that the number of in-links in the test collection is significantly lower than the number of out-links, which degrades the effectiveness of the Co-citations measure. Then, the authors have investigated different strategies for combining citation-based classifiers using

Amsler similarity measure with content-based classifiers. However, they found that the gains obtained by each combination method are quite small as compared to the best performing classifier.

Several hybrid studies (Aljaber, Martinez, Stokes, & Bailey 2011; Cao & Gao 2005; B. Zhang et al. 2005) are suggested to classify papers by exploiting content-based and citations-based information. For instance, in (B. Zhang et al. 2005), the authors apply a Genetic Programming technique that fuse different kinds of content and bibliometric similarity measures to obtain better similarity functions. Their experiments show that the fused similarity functions, which are used by KNN classifiers, lead to better results than the best similarity functions.

In (Cao & Gao 2005), the authors have suggested a framework that incorporates direct citation information in the form of a symmetric adjacency matrix, wherein each cell indicates whether a direct citation link (i.e. in-link and/or out-link) exists between a pair of papers, and content information in the form of a document-term matrix. Then, a joint factorization of the two matrices is performed to produce a compact data representation in a lower dimensional space. Their results show that the factored data representation leads to better results of the SVM classifier with regards to the data representation using content-based or citation-based features.

In (Aljaber et al. 2011), the authors suggest an approach for the classification of biomedical papers with Mesh terms by exploiting their content and citation contexts information. Citation contexts, which correspond to the short texts surrounding citations markers that refer to a paper in other publications, are found to be rich sources of semantically similar words (e.g. synonyms, hypernyms) that do not appear in the original content of the cited paper. In their approach, citation contexts referring to a paper are exploited in conjunction with its content information (i.e. full text) to obtain a richer BoW representation. Obtained results have confirmed the advantage of using the enriched BoW representation on the performance scores. Besides, the best performances are obtained when using only terms from citation contexts holding a semantic relationship (i.e. synonymy, hypernymy) with a word from the original paper's content, in conjunction with their synonyms.

The main disadvantage of the aforementioned supervised approaches is that they assume the presence of huge amounts of labeled training data, which are necessary for training a supervised classifier. However, in real world situations, obtaining such collections is a difficult and time-consuming task.

2) Semi-supervised approaches: These studies tackle the scarcity issue of labeled data of the use of unlabeled data. In (Caragea, Bulgarov, & Mihalcea 2015), a Co-training-based approach is suggested for the single-label classification task of research papers while exploiting papers' contents (i.e. title, and abstract parts) and papers' citation contexts as two distinct views. Differently from (Aljaber et al. 2011), the authors consider two kinds of citation contexts for a paper: cited contexts, which are the short texts surrounding its citation mention in other publications, and citing contexts which correspond to the sentences surrounding the citation markers present in its content referring to its cited references. In their work, two feature vector representations are generated for each paper based on its content and citation contexts information (i.e. citing and cited contexts). The latter are separately exploited by two Co-training classifiers, which iteratively add their reliable predictions on unlabeled samples to the labeled set. Unlike the original Co-training algorithm, the most confidently predicted unlabeled samples by each classifier are those predicted with a probability score surpassing a fixed threshold. Their experiments validate the efficiency of their suggested Co-training framework in improving the performance scores by the exploitation of unlabeled data, and prove the advantage of using both types of citation contexts.

In (Laguna & de Andrade Lopes 2009), the authors suggest a Co-training-based approach for the single-label classification task of research papers while exploiting content and citation-based information, as two distinct views. In their approach, the content-based classifier is learnt using the BoW representation of the papers' content, while the citation-based classifier is trained using a feature representation derived from a symmetric adjacency matrix capturing direct relationships between papers. The content-based and citation-based classifiers are iteratively learnt using both labeled and unlabeled data, similar to the original Co-training algorithm. Their experimental evaluation proves the efficiency of their suggested Co-training-based framework in achieving better results.

Further to prior conducted review, we deduce that very few semi-supervised studies (Caragea et al. 2015; Laguna & de Andrade Lopes 2009) are suggested to classify papers using limited amounts of labeled data with a larger set of unlabeled data. However, suggested studies target the single-label classification task, and thus are not applicable in HMC settings.

Differently from existing researches, the present work suggests a novel semi-supervised approach for the HMC task of scientific papers that takes advantage of unlabeled data to achieve higher classification results. The proposed approach, which relies on the Co-training paradigm, uses content and bibliographic coupling information as two distinct views of a scientific paper. Specifically, in our approach, each paper is described by two feature vector representations corresponding to its content and bibliographic coupling-based views, respectively. The first corresponds to a BoW representation of the paper's content (i.e. abstract and title sections). The second is derived from a similarity matrix wherein the pairwise papers similarities are computed using a novel bibliographic coupling measure, called DescriptiveBC (Liu 2017). The latter does not only search for the common out-link citations between papers, but also takes into account the similarity between out-link citations titles. Based on two feature vector representations of data, two hierarchical multi-label classifiers are separately learnt on different views of the labeled data, and iteratively select their most confidently predicted unlabeled samples using a novel selection mechanism.

3 | PROPOSED APPROACH

In this section, we present our proposed Co-training-based approach for the HMC task of research papers, which exploits content and bibliographic coupling information as two distinct papers' views. In the following subsections, we firstly introduce some basic notations. Secondly, we present the used hierarchical multi-label classifier. Thirdly, we give details about MLSMOTE method (Charte et al. 2015). Fourth, we explain the suggested criteria for confident samples selection. Finally, we present the complete algorithm.

3.1 | Notations

In a typical Co-training framework, we have an input feature space $X = X^{(1)} \times X^{(2)}$ where $X^{(1)}$ and $X^{(2)}$ correspond to two independent and sufficient views of a paper, i.e. content-based and bibliographic coupling-based views, respectively. Let $Y = \{0,1\}^q$ a q dimensional label space, where $q = |Q|$ and $Q = \{l_1, l_2, \ldots, l_q\}$ is the set of labels that are arranged in a hierarchy. Each paper i is represented by a feature vector $x_i = (x_i^{(1)}, x_i^{(2)})$, where $x_i^{(1)}$ is the content-based feature vector, and $x_i^{(2)}$ is the bibliographic coupling-based feature vector. Each paper can be equally associated to a labels vector $y_i = [y_{i1}, y_{i2}, \ldots, y_{iq}] \in Y$, where $y_{ij} = 1$ if label l_j is assigned to x_i, $y_{ij} = 0$ otherwise (i.e. l_j is a negative label). Assume we have a small labeled dataset $L = \{(x_i, y_i)\}_{(i=1)}^{l}$ containing l labeled samples and a larger unlabeled dataset $U = \{x_i\}_{i=l+1}^{l+u}$ of u samples with unknown labels. The objective is to perform an efficient Co-training learning of two hierarchical multi-label classifiers $C^{(1)} : X^{(1)} \rightarrow Y$, and $C^{(2)} : X^{(2)} \rightarrow Y$, which exploit both L and U training datasets to achieve better performance results on unseen samples.

3.2 | Hierarchical multi-label classifier

The hierarchical multi-label classifier, used in this work, follows an LCN-based local approach that uses multiple binary classifiers for each label/node in the hierarchy. At the prediction step, each binary classifier provides its decision concerning the relevancy of a specific label to an unlabeled sample. Then, the local classifiers outputs are combined to deduce a preliminary set of assigned labels, which are not necessarily complying with the hierarchy constraint. The final labels are obtained by applying a bottom-up rule that propagates the positive predictions from lower to higher nodes in the hierarchy.

In this work, we use Random Forest (Breiman 2001) as the basic binary classifier because it is one of the rules-based learners that are known by their good capacities in handling intuitively concept drifts. Specifically, rules-based classifiers are often used to deal with concept drifts, which refer to the changes of the underlying distribution of unlabeled data over the time, that are usually occurring in evolving learning data situations (S. Wang, Minku, & Yao 2018).

3.3 | MLSMOTE

In HMC settings, the labeled dataset is characterized by a desequilibrated labels distribution. As a result, classifiers face difficulty in discovering the true labels of unlabeled samples since they are biased towards the majority labels, and cannot detect the minority ones. To tackle this issue, we use MLSMOTE (Charte et al. 2015), which is one of the most popular oversampling algorithms for multi-label data, to rebalance the labels distribution of the initial labeled set L. Specifically, MLSMOTE generates synthetic samples for the minority labels in the initial labeled dataset. The first step in MLSMOTE consists in discovering the minority labels in the initial labeled dataset L using two imbalance level measures: IRlbl and MeanIR (Charte, Rivera, del Jesus, & Herrera 2013). More precisely, the IRlbl measure estimates the imbalance ratio for a given label in dataset L while the MeanIR measure computes the average imbalance level in L. The labels, whose IRlbl values are greater than the MeanIR score, are considered as the set of minority labels. Then, the minority samples in the dataset L (i.e. samples associated with one of the minority labels) are selected, and used as seeds for the generation of new synthetic instances. More precisely, for each minority sample (x_i, y_i), a new synthetic instance is created as follows: the k minority nearest neighbors to (x_i, y_i) are selected from the labeled training dataset L. Then, one of the neighbors is randomly chosen. After that, the feature vector of the synthetic sample is created, for numeric attributes, by interpolation of the features values belonging to the minority sample and its random neighbor. As for nominal attributes, their values are assigned using a majority voting method that considers the features values of the neighboring samples.

Finally, the label set of the synthetic sample is deduced using a ranking method that chooses the labels, which are present in half or more of the neighbors. Following the described process, synthetic samples are generated for all the minority labels based on the original minority samples. As output, MLSMOTE returns the rebalanced dataset with the newly generated synthetic samples.

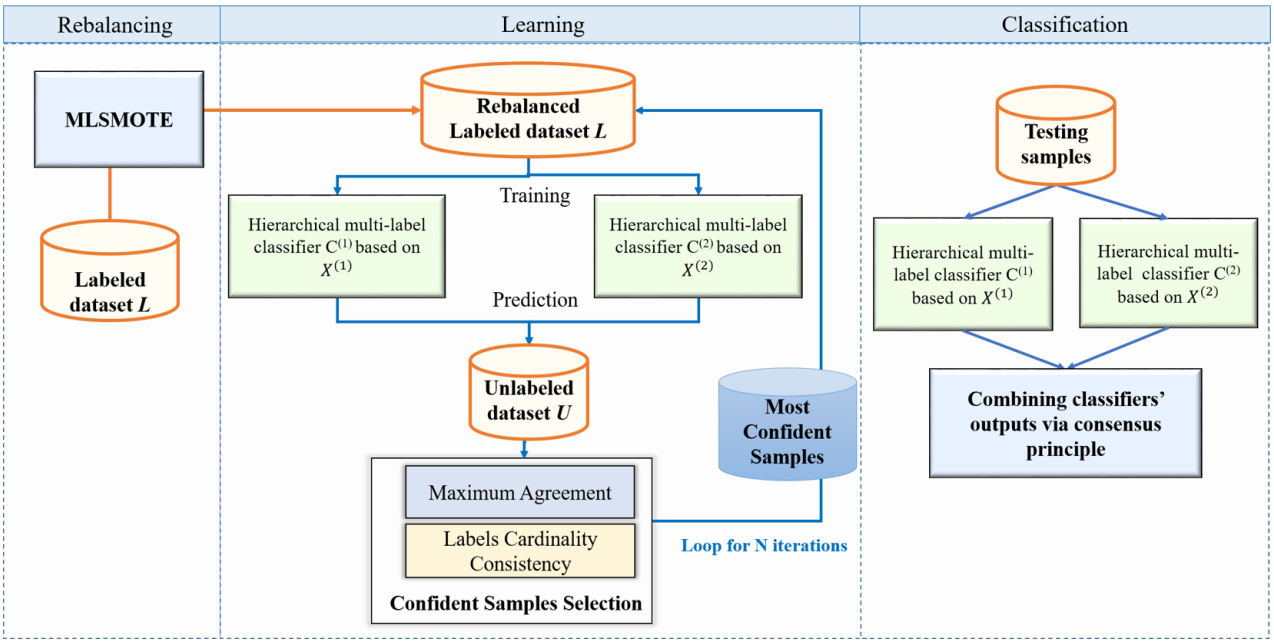


FIGURE 1 Overview of the proposed approach.

3.4 | Suggested Criteria for Confident Samples Selection

Deciding which unlabeled samples are associated with confident labels is necessary to ensure a robust Co-training learning. The standard Co-training algorithm (Blum & Mitchell 1998) chooses the unlabeled samples having the highest confidence score of the most likely label. Thus, it is not appropriate for the HMC task wherein samples are associated with many labels simultaneously. In HMC settings, the main challenge is how to check the reliability of all generated labels predictions for an unlabeled sample. An efficient selection strategy should avoid the addition of misclassified unlabeled samples, i.e. samples having noisy assigned labels and/or with missing relevant labels. Thus, to prevent the selection of such misclassified samples, we define two selection criteria which are: Maximum Agreement and Labels Cardinality Consistency that allow discovering the most confidently predicted unlabeled samples by each of the hierarchical multi-label classifiers $C^{(1)}$ and $C^{(2)}$.

The Maximum Agreement criterion selects the unlabeled samples on which the two classifiers exploiting different views have a higher consensus on their labels predictions, which guarantees a better reliability of their assigned labels. The Labels Cardinality Consistency criterion discards the unlabeled samples having an inconsistent labels cardinality by considering the number of assigned labels to the original labeled samples.

Note that during a Co-training round, only the classifier $C^{(j)}$ under the j^{th} view ($j \in \{1, 2\}$) is considered active, i.e. devoted to assign labels for unlabeled samples. In the following, we explain how each criterion determines the most reliable predictions made by the active classifier $C^{(j)}$ in a Co-training round.

3.4.1 | Maximum Agreement Criterion

Identifying the well-predicted unlabeled samples by the active classifier $C^{(j)}$ in a Co-training round can be performed by exploiting the agreement information on the labels prediction among the two classifiers, exploiting different views (Sousa & Gama 2017). In other words, we assume that when classifiers $C^{(1)}$ and $C^{(2)}$ maximally agree on the predicted relevant and irrelevant labels for an unlabeled sample, the latter is more likely to be properly labeled. Thus, we seek to select the unlabeled samples with the highest agreement levels on labels predictions, which tend to be the most confident ones. Given an unlabeled sample $(x_i, y_i) = C^{(j)}(x_i^{(1)})$ and $z_i = C^{(2)}(x_i^{(2)})$ are two predicted labels vectors for x_i by classifiers $C^{(1)}$ and $C^{(2)}$, respectively. The agreement level for x_i is computed as the sum of the common labels predictions in y_i and z_i, as defined by the following equation:

$$Arg(x_i) = \sum_{k=1}^{q} 1_{(y_{ik}==z_{ik})} \quad (1)$$

Where q is the number of labels.

Using equation (1), samples with the highest agreement levels can be discovered by maximizing the value of $Agr(x_i)$.

3.4.2 | Labels Cardinality Consistency criterion

Maximum agreement criterion helps identifying the well-predicted unlabeled samples, but cannot really check the labeling mistakes made by the currently active classifier $C^{(j)}$ in a Co-training round. In this work, we consider the number of predicted relevant labels to unlabeled samples as an indicator of the classification errors. The main idea here is that labeled and unlabeled data share similar statistical properties in the label space (Li & Guo 2013; Wei, Wang, & Zhao 2013). For that, we assume that the cardinality of the predicted relevant labels for an unlabeled sample should be consistent with the number of associated labels to the original labeled samples. Given a candidate unlabeled sample $(x_i, y_i) = C^{(j)}(x_i^{(j)})$ is its predicted labels vector by the active classifier $C^{(j)}$ in a Co-training round. The consistency of the cardinality of the assigned labels to x_i is checked using the following equation:

$$\min_{(x'_i,y'_i)\in L'} |y'_i| \leq |y_i| \leq \max_{(x'_i,y'_i)\in L'} |y'_i| \tag{2}$$

Where $|y_i|$ denotes the number of the predicted relevant labels in y_i, and L' is the original labeled set.

Based on (2), we specify that the number of predicted relevant labels for an unlabeled sample x_i should not be higher than the maximum number of associated labels to the original labeled samples L'. It should not be also lower than the minimum number of assigned labels to the original labeled samples. Every unlabeled example that does not meet these constraints has more chances to be misclassified, and thus should be excluded from the set of candidate confident unlabeled samples.

3.5 | Proposed Algorithm

Algorithm 1 describes the details of the proposed Co-training-based framework for the HMC task of scientific papers. Firstly, the initial labeled set L is enriched with new minority synthetic samples using the MLSMOTE(L, k) function. This function takes as inputs: the labeled dataset L to be oversampled and the number of nearest neighbors k (line 10). It returns the rebalanced dataset with newly generated minority synthetic samples by MLSMOTE method (Charte et al. 2015), as described in section 3.1.2. Then, two hierarchical multi-label classifiers $C^{(1)}$ and $C^{(2)}$ are trained on the rebalanced labeled set using distinct features sets, corresponding to different papers' views, and are iteratively refined by the exploitation of the confident predictions of each other (lines 12-33). In each Co-training iteration, one of the classifiers $C^{(1)}(j \in \{1, 2\})$ is considered active, and its m most confidently predicted unlabeled samples are selected through an iterative loop (lines 13-32). Firstly, for each unlabeled sample x_i in the unlabeled set U, the agreement level on its predicted relevant and irrelevant labels among the classifiers $C^{(1)}$ and $C^{(2)}$ is computed, as defined in equation (1). Then, the unlabeled sample $\check{x}_i$ having the highest agreement level is picked (line 17). After that, its labels vector $\check{y}_i$ is predicted by the active learner $C^{(j)}$, to check the consistency of the cardinality of its assigned labels (line 23). Once the required constraints in equation (2) are satisfied, $\check{x}_i$ is assumed to be associated with a consistent labels number. In this case, $\check{x}_i$ is added to the pool of candidate confident samples $\hat{U}$ (line 24). The described process repeats until $\hat{U}$ contains the m samples with the maximum agreement levels, and assigned with a consistent number of assigned labels by the currently active classifier $C^{(j)}$. The latter are considered as the most confidently predicted samples by the active classifier $C^{(j)}$, which are added (with their predicted labels) to the labeled set, and removed from the unlabeled set (lines 27-30). The iterative Co-training learning continues until the maximum number of iterations N is reached. At the end, we get two final classifiers $C^{(1)}$ and $C^{(2)}$, which are used during the classification step for inducing the labels of unseen test samples. Specifically, the label sets of unseen samples are assigned based on the consensus principle, i.e. a label is assigned to an unseen test sample only when it is considered relevant by both classifiers.

4 | EVALUATION STUDY

In this section, we describe the established experimental framework for the evaluation of our approach compared to several baselines. We firstly give a description of the experimental dataset. After that, we give a description of the experimental setup, and the used evaluation measures. Finally, we discuss the obtained results on the realized experiments.

4.1 | Experimental Dataset

4.1.1 | Dataset Acquisition

To assess the performance of our approach, we use a collection of 3170 scientific papers retrieved from the ACM digital library [2], wherein papers are annotated according to the predefined labels in the ACM classification tree. The papers were automatically extracted using the web data extraction

[2] https://dl.acm.org/

tool Connotate. Specifically, using Connotate [3], we create different agents that are initially configured with an initial address (i.e. an URL in the ACM library) to crawl their linked paper pages. In fact, the agents are trained in a way that they crawl only the pages pointing to a scientific paper to extract its different metadata (i.e. title, abstract, cited references, and its assigned labels from the ACM tree). Since some papers have some missing metadata, we specified that our collected ACM dataset contains only the papers having non-empty values for these metadata. Moreover, because of the large size of the current version of the ACM classification tree, we considered only the labels appearing in the first and second hierarchical levels. We also restricted that considered labels in the classification task should have at least 120 representative examples in the whole dataset.

Algorithm 1 PROPOSED ALGORITHM

INPUT
L : labeled set
U: unlabeled set
$X^{(1)}, X^{(2)}$: two feature sets corresponding to different views
k: number of nearest neighbors used by MLSMOTE
m: the final batch size of confident samples
N: the maximum number of iterations
Begin
$L' = L$
L= MLSMOTE (L, k)
iterations=0
repeat
  for $j \in \{1, 2\}$ **do**
    Train hierarchical multi-label classifier $C^{(1)}$ using L based on $X^{(1)}$
    Train hierarchical multi-label classifier $C^{(2)}$ using L based on $X^{(2)}$
    for each x_i in U **do**
      Compute $Agr(x_i)$ using equation (1)
    end for
    $\hat{U} = \emptyset$
    while $(|\hat{U}| \leq m)$ **do**
      $\check{x}_i = \underset{x_i \in U}{\text{argmax}}(x_i)$
      $\check{y}_i = C^{(j)}(\check{x}_i^{(j)})$
      if $\min\limits_{(x'_i, y'_i) \in L'} |y'_i| \leq |y_i| \leq \max\limits_{(x'_i, y'_i) \in L'} |y'_i|$ **then** ▷ check whether the defined constraints in equation (2) are satisfied
        $\hat{U} = \hat{U} + \check{x}_i$
      end if
    end while
    for each $\check{x}_i$ in $\hat{U}$ **do**
      $\check{y}_i = C^{(j)}(\check{x}_i^{(j)})$
      Add the unlabled sample $(\check{x}_i, \check{y}_i)$ to L ($L \leftarrow L + \check{x}_i, \check{y}_i$) and remove it from U ($U \leftarrow U - \check{x}_i$)
    end for
    iterations= iterations+1
  end for
until (iterations $\geq$ N)
OUTPUT: Two final classifiers $C^{(1)}$ and $C^{(2)}$ used to discover the labels on each unseen sample based on the consensus principle.

[3] https://www.connotate.com

4.1.2 | Feature vectors generation

In this work, two distinct feature vector representations are generated for each paper from the collected raw dataset corresponding to its content-based and bibliographic coupling-based views. In the following, we detail the feature vectors generation principle.

- **Content-based feature vector**

We build a content-based feature vector for each paper from the raw dataset based on the occurrences of the vocabulary terms in that paper. The first step is to pre-process the input raw dataset to choose the vocabulary terms from the textual contents of all papers (i.e. their abstract and title sections). The pre-processing step includes some NLP tasks such as tokenization and Part Of Speech (POS) tagging, which decompose the papers' textual contents into a set of words with their associated POS tags [4]. We keep only the terms having their POS tags as nouns, adjectives, verbs or adverbs, which facilitates the removal of all the unnecessary terms (e.g. conjunction and stop words). The selected words are lowercased, and then stemmed using the Porter Stemmer algorithm (Porter 2006) in order to reduce the amount of redundant features. After that, pruning is carried out to eliminate the low-frequency terms, i.e. the terms occurring less than three times in different papers from the dataset. The remained words constitute the resulting vocabulary terms. After selecting the vocabulary terms, a paper-term matrix P is constructed where rows correspond to papers from the input dataset, columns refer to the vocabulary terms, and cells are the terms weight values in each paper. We adopt the Term-Frequency (TF) weighting schema (Salton & Buckley 1988) that simply counts the occurrence of each term in the considered paper. The terms weight values are further scaled to the range $[0, 1]$ based on the min-max normalization principle that considers the minimum and maximum values of the terms occurrences in each paper from the whole dataset (Jayalakshmi & Santhakumaran 2011). Thus, the cell $P_{(i,j)}$ denotes the normalized TF value of the j^{th} term in the i^{th} paper. In the matrix P, the i^{th} row corresponds to the BoW representation of the i^{th} paper in the dataset.

- **Bibliographic coupling-based feature vector**

To quantify the citation links strength between two papers, we use an enhanced bibliographic coupling measure, called DescriptiveBC (Liu 2017). The main advantage of DescriptiveBC with respect to the original BC measure is that it does not only search for the common cited references (i.e. common out-link citations), but also takes into consideration the similarity between the references' titles, which allows to tackle the scarcity issue of common out-link citations between papers in real datasets. Thus, papers having similar titles of cited references (i.e. sharing common words) are considered as highly related. Specifically, using DescriptiveBC, the relatedness between two papers i and j is computed based on the similarity between their references' titles by selecting for each reference, in paper i, its most similar reference in paper j, and vice versa. Then, the computed individual similarities are summed, and normalized by the total number of out-link citations in both papers, as defined by the following equation:

$$DescriptiveBC(i,j) = \frac{\sum_{r_1 \in R_i} Max_{r_2 \in R_j} Sim(r_1, r_2) + \sum_{r_2 \in R_j} Max_{r_1 \in R_i} Sim(r_1, r_2)}{|R_i| + |R_j|} \tag{3}$$

Where $|R_i|$ and $|R_j|$ denote the number of out-link citations in papers i and j, respectively, and $\mathrm{Sim}(r_1, r_2)$ is the similarity value between two references r_1 and r_2 cited by papers i and j respectively, as defined in equation (4). $\mathrm{Sim}(r_1, r_2)$ equals to 1 when references r_1 and r_2 are identical. Otherwise, $\mathrm{Sim}(r_1, r_2)$ is computed using Jaccard index measure while considering the reference terms as two distinct sets. The latter computes the similarity between two references r_1 and r_2 as the number of intersecting terms divided by the cardinality of all their terms.

$$Sim(r_1, r_2) = \begin{cases} 1 & \text{if } r_1 = r_2 \\ \frac{|\mathrm{Title}(r_1) \cap \mathrm{Title}(r_2)|}{|\mathrm{Title}(r_1) \cup \mathrm{Title}(r_2)|} & \text{otherwise.} \end{cases} \tag{4}$$

Where $\mathrm{Title}(r_1)$ represents the set of terms in reference r_1.

Given the pairwise similarities between papers from the input dataset, we construct a similarity matrix M, wherein each cell M_{ij} holds the approximated DescriptiveBC similarity between two papers i and j. Thus, the constructed vector M_i represents the bibliographic coupling-based feature vector of paper i that contains its similarity scores towards every paper j in the dataset.

4.1.3 | Dataset Description

The constructed dataset is described by a set of characteristics: the number of instances (I), features (F), labels (L), distinct label sets (DL), the density (D), the cardinality (C), the imbalance ratio (IR), and the depth of the hierarchy (Dh), which are summarized in Table. I. The imbalance ratio

[4] Tokenization and Part Of Speech (POS) tagging tasks are performed using the Stanford CoreNLP tool https://stanfordnlp.github.io/CoreNLP/

TABLE 1 Description of the Experimental Dataset.

DATASET	I	F	L	D	DL	C	DH	IR
ACM	3170	7617	16	0.14	151	2.32	2	2.37

(IR) is computed using the MeanIR measure (Charte et al. 2013). As for the cardinality of a dataset, it corresponds to the average of the number of labels of all the dataset instances (Tsoumakas & Katakis 2007). Let DS a dataset that contains N labeled examples.

$$Cardinality(DS) = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{5}$$

The density of the dataset D corresponds to the mean of the number of labels of its samples divided by the total number of labels, Q.

$$Density(DS) = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{Q} \tag{6}$$

From Table I., we see that our experimental ACM dataset contains only 16 labels, but a high number of distinct labels combination (i.e. DL=151), which makes it difficult to learn by machine learning classifiers.

4.2 | Experimental Setup

To evaluate the performances of the different compared methods, we divide each dataset into four folds: one for the test phase and the three other folds for the training step. Then, the training dataset is partionned into two sets; one represents the labeled dataset while the other corresponds to the unlabeled dataset. Since we want to evaluate the impact of the sizes of both labeled and unlabeled datasets on the classification performances, we have used varied ratios of labeled examples number (Ln) to unlabeled samples number (Un) in all conducted experiments, i.e. (Ln:Un)=(1:2), (1:3), and (1:4). Note that we apply stratified sampling to divide the dataset into equal parts with equilibrated labels distribution. This guarantees that classifiers are learnt, and assessed on all class labels.

All Performed experiments are repeated 10 times with different random folds, and then obtained results are averaged to get the final scores. In our experiments, the different input parameters are fixed as follows: N=80, k=5 (the default value in MLSMOTE implementation[5]), and m=10. We use the implementations of the Random Forest classifier provided by the Mulan package[6], which is an open source Java library for multi-label learning. Experiments are performed on a PC server with Processor Intel Xeon(R), 2.20 GHz x32, 64 GB RAM.

4.3 | Evaluation Measures

In this work, we choose to use two evaluation measures, which are: Micro-F1 and Macro-F1 (Wu & Zhou 2017). Given C a hierarchical multi-label classifier, y_i is the hierarchical multi-label prediction for an instance x_i by classifier C, z_i is the true label set of x_i, N denotes the number of instances in the test dataset, and Q refers to the number of labels.

- Micro-F1: is calculated as the harmonic mean of Micro_recall and Micro_precision, as defined in (7). This measure combines the predictions of all the labels, and then estimates the overall classification performance over all sample/label pairs. Thus, it is implicitly giving the opportunity to the majority labels to influence the scores.

$$Micro - F1 = 2 \times \frac{Micro\_Precision \times Micro\_Recall}{Micro\_Precision + Micro\_Recall} \tag{7}$$

Where Micro_precision and Micro_recall are defined as follows:

$$Micro\_precision = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{q=1}^{L} y_{iq} z_{iq}}{\sum_{q=1}^{L} z_{iq}} \tag{8}$$

$$Micro\_Recall = \frac{1}{N}\sum_{i=1}^{N} \frac{\sum_{q=1}^{L} y_{iq} z_{iq}}{\sum_{q=1}^{L} y_{iq}} \tag{9}$$

[5]The Java application is available at: https://simidat.ujaen.es/řesearch/MLSMOTE/index.html
[6]http://mulan.sourceforge.net/

- Macro-F1: computes the precision and recall values for each label k independently, and then averages across all labels. By this way, all labels have equal weights, which gives more emphasis to the minority labels to influence the results.

$$Macro - F1 = \frac{1}{Q}\sum_{k=1}^{Q}\frac{Precision_k \times Recall_k}{Precision_k + Recall_k} \tag{10}$$

Where Precision_k and Recall_k score values are computed by the following equations:

$$Precision_k = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_{ik} \cap y_{ik}|}{|z_{ik}|} \tag{11}$$

$$Precision_k = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_{ik} \cap y_{ik}|}{|y_{ik}|} \tag{12}$$

4.4 | Experiments' Description, Results and Discussion

In this section, we evaluate the efficiency of the proposed approach and its different components under different scenarios. The objectives of the conducted experiments can be formulated by several research questions:

- **Is our proposed Co-training-based approach a successful semi-supervised learning method?**

The goal of the first experiment is twofold. First, we show the power of our semi-supervised approach in benefiting from unlabeled examples for achieving better classification results. Second, we demonstrate the usefulness of combining the two suggested selection criteria, i.e. Maximum Agreement and Labels Cardinality Consistency, to sustain a successful Co-training learning. For that, we compare the performance of the proposed approach with the following baseline methods:

– **Supervised baseline:** It uses two hierarchical multi-label classifiers, which are learnt on the rebalanced labeled set, using each the features corresponding to one of the papers' views. The classifiers outputs are combined via the consensus principle used in our proposed approach.

– **Maximum Agreement (MA):** A variant of the proposed Co-training-based approach that considers only the Maximum Agreement criterion in confident samples selection. It chooses the m samples having the maximum agreement levels as the most confidently predicted ones.

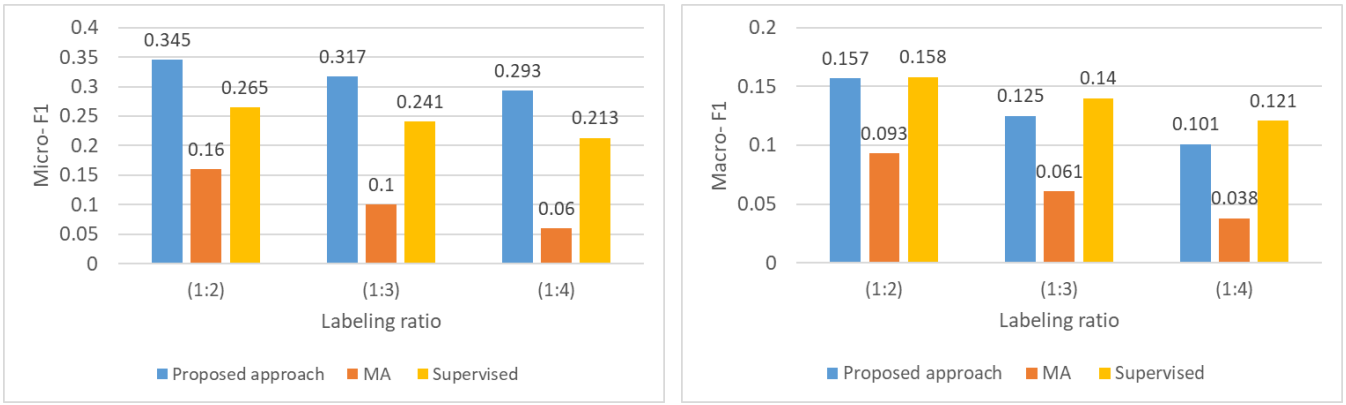
Figure 2 presents the obtained results for the comparative methods on our experimental dataset at different labeling rates in terms of Micro-F1 and Macro-F1 scores. Many general conclusions can be made upon these results. First, we observe that our proposed approach significantly outperforms the supervised baseline in terms of Micro-F1 scores at varied labeling ratios. This indicates that, in our approach, the ratio of the relevant labels predictions (i.e. trues positives) to the number of irrelevant ones (i.e. false positives, and false negatives) is enhanced. These results prove the success of the proposed approach in benefiting from the newly labeled examples for enhancing the overall performance scores on all samples, and the efficiency of our confident selection mechanism.

Second, we found that, as expected, when the size of the labeled training data increases, the obtained performance scores of our proposed approach are better in terms of Micro-F1 scores. However, the best improvement achieved by our approach over the supervised baseline in terms of Micro-F1 scores occurs at the labeling ratio (Ln:Un)=(1:4). These results reveal that our proposed approach handles well the case when few training samples are available, similarly to several semi-supervised studies (e.g. (Q. Wang, Li, & Gool 2019)).

Third, we observe that, under all labeling ratios, our approach leads to lower Macro-F1 scores with regards to the supervised baseline. Since the Macro-F1 measure gives more importance for labels with low frequencies to influence the scores, we can deduce that the individual performances on the minority labels has decreased, indicating an accumulation of the misclassification errors for these labels. In other words, the ratio of irrelevant predictions to the amount of correct predictions for the minority labels has increased, which results in the decrease of the Macro-F1 scores. From these results, we deduce that, in our semi-supervised approach, majority labels are better-predicted, leading to an increase of the Micro-F1 scores, contrarly to the minority ones.

This phenomenon is explainable. First, the obtained performance scores by our approach are significantly higher for the Micro-F1 scores than for the Macro-F1 scores, which are very low, at the beginning of the Co-training learning (i.e. iteration 0). Thus, initial classifiers have very weak capacities in predicting the minority labels, but have relatively strong capabilities in predicting the majority ones.

Consequently, the most confident unlabeled samples chosen in each iteration, which are those having the highest agreement levels on their labels predictions among hierarchical multi-label classifiers, would be often assigned with majority labels that tend to be correct, especially at the first learning rounds. Thus, classifiers become more robust on predicting the majority labels after adding the confident unlabeled samples, which enhances the classifiers' performance on these labels, which leads to an improvement of the Micro-F1 scores.



(a) Micro-F1 scores

(b) Macro-F1 scores

FIGURE 2 Classification performances of the proposed approach, MA and the supervised baseline in terms of both evaluation measures. The horizontal axis represents the different labeling ratios while the vertical axis indicates the used evaluation measure.

However, added samples have more chances to include irrelevant predictions for the minority labels (i.e. a minority label is wrongly assigned or not detected), which weakens the behavior of the classifiers on these rare labels. Consequently, with the progress of the Co-training learning and the addition of more unlabeled samples, the number of misclassifications (false positives and false negatives) for the minority labels would increase, which results in the decrease of the Macro-F1 scores. This observation explains why our semi-supervised approach yields inferior Macro-F1 scores at all labeling ratios as compared to the supervised baseline.

Finally, we observe that, with different labeling ratios, the proposed approach consistently outperforms the MA method in terms of Micro-F1 and Macro-F1 scores. We also found that, under all cases, MA method has failed to improve the performance scores compared to the purely supervised method. This indicates that, in MA method, there is an accumulation of the classification errors, because of the addition of too many misclassified unlabeled samples, which leads to a deterioration of the performance scores with regards to the purely supervised baseline. Consequently, these results demonstrate the necessity of combining the two suggested criteria to sustain a beneficial Co-training learning. Hence, we can conclude that an appropriate selection mechanism is critical for the success of the Co-training learning in HMC settings.

- **What is the impact of the oversampling method on the Co-training learning?**

In this experiment, we prove that, by exploiting the synthetic samples, which are generated by the MLSMOTE oversampling method, the Co-training learning is performed in a better-balanced manner. For that, we compare the performance results of our Co-training-based approach against another variant of our approach that do not integrate MLSMOTE, which is denoted as Co-Training Without Oversampling (CTWO) method. Comparisons are made to discover the influence of the inclusion of the oversampling technique at varied labeling ratios, and at different learning iterations. Figure 3 shows the performance results of both methods on our experimental dataset under different labeling ratios in terms of Micro-F1 and Macro-F1 scores. Several conclusions can be drawn upon this figure.

First, we observe that CTWO method achieves the highest performance scores in terms of Micro-F1 and Macro-F1 scores at the beginning of the Co-training learning (i.e. iteration 0). Contrarily, our proposed method, which takes advantage of synthetic data, usually leads to higher performances than CTWO method in terms of both measures at the end of the Co-training learning (i.e. iteration 80). This indicates that the proposed approach better handles the label imbalance issue during the Co-training learning, which results in higher final performance scores.

Second, we find that CTWO method always achieves the best results in terms of Micro-F1 scores during the first iterations. This is because, in CTWO method, classifiers well predict the majority labels, and tend to misclassify the minority ones. Thus, the selected confident data by CTWO method would be generally associated with the majority labels, which makes it possible to maintain high improvement rates in terms of Micro-F1 scores at the first learning iterations. For example, at the sample rate (Ln:Un) = (1:2), wherein the size of the labeled training data is the largest, the CTWO method achieves the best performances during the first 50 iterations.

Thirdly, we notice that, with the progress of the learning process, the two compared methods suffer from the imbalance problem, which has a negative influence on the performance scores. However, the behavior of the two compared methods differs significantly with varied ratios. At the sampling rate (Ln: Un) = (1: 2), we find that CTWO method achieves progressive improvements in terms of Micro-F1 scores during the first 50 iterations. However, it gets the optimal performances in terms of Macro-F1 scores at iteration 20, before degrading progressively and slightly during the following iterations. This indicates that, from iteration 20 to iteration 50, the Micro-F1 scores are increased while the Macro-F1 scores are degraded. From these results, we deduce that the added confident samples by CTWO are mostly associated with the majority labels

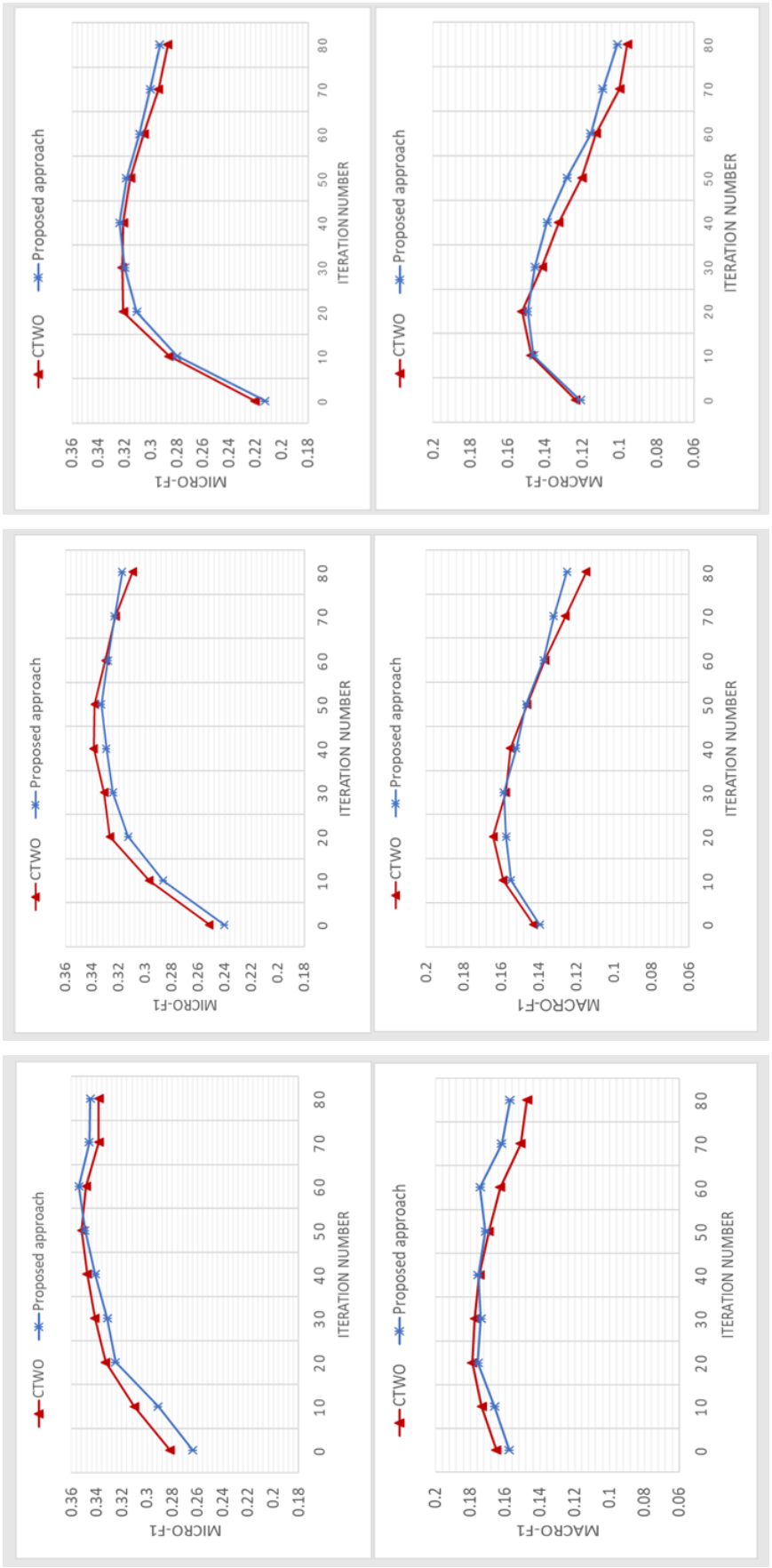


FIGURE 3 Performance results of the proposed approach and CTWO (Co-training without oversampling) method across different learning iterations on our experimental dataset at different labeling ratios. From top to bottom, we present the evaluation figures based on Micro-F1 and Macro-F1 scores, respectively, at each labeling ratio.

since classifiers become increasingly well predicting the majority labels to the detriment of the minority ones. Thus, with the exaggeration of the imbalance problem, the misclassification errors are accumulated, leading to lower results in terms of Micro-F1 scores at the later iterations, i.e. from iteration 60 to iteration 80.

Unlike CTWO, our approach leads to a more balanced learning of the different labels. As an example, from iteration 40 to iteration 50, we find that the proposed approach achieves a more significant improvement rate in terms of Micro-F1 scores, as compared to CTWO method. In terms of Macro-F1 scores, our approach obtains a slightly lower degradation rate. These results indicate that the misclassification errors, which are exaggerated by the imbalance problem, are minimized in our proposed approach. For that, it reaches the peak Micro-F1 score at a later iteration of the learning process, i.e. iteration 60, and maintains significant improvements in terms of both evaluation measures at further iterations.

At the sampling rate (Ln: Un) = (1: 3), we observe that CTWO method encounters progressive improvements in terms of Micro-F1 scores during the first 40 iterations. However, in terms of Macro-F1 scores, it achieves its optimal performances at iteration 20, before degrading at further iterations. As it can be seen, from iteration 20 to iteration 30, CTWO method obtains a less significant improvement in terms of Micro-F1 scores while its Macro-F1 scores are significantly decreased. These results show that the imbalance problem is reinforced since classifiers have become increasly overwhelmed by the majority labels to the detriment of the minority ones. Thus, by adding more majority samples, the imbalance problem is exaggerated, which increases the misclassification rates significantly. As a result, the performance scores in terms of Micro-F1 are progressively declined during the last 40 iterations.

Unlike CTWO, we find that the imbalance problem, which is pronounced beginning from iteration 30, is less significant in our proposed approach. For example, from iteration 20 to iteration 30, our approach achieves a more significant improvement in terms of Micro-F1 scores as compared to CTWO method while its Macro-F1 scores are increased. These results indicate that, in our approach, classifiers are less biased towards the majority labels, which decreases the misclassification errors during the following iterations. This explains why our approach was able to obtain its best Micro-F1 score at a later iteration, i.e. iteration 50, and to reach superior results at the end of the Co-training learning.

At the sampling rate (Ln:Un) = (1:4), we find that CTWO method maintains progressive improvements in terms of Micro-F1 scores only during the first 30 iterations. Similar to previous cases, the obtained Macro-F1 scores by CTWO method starts to degrade beginning from iteration 20. However, in this case, the degradation rates of the Macro-F1 scores are more significant. This can be explained by the fact that, in this case, the size of the labeled training set is smaller, thus discriminating the majority labels from the minority ones is more difficult. Moreover, the number of training samples for the minority labels is smaller, which increases the classification errors. For example, from iteration 20 to iteration 30, the Micro-F1 scores are slightly enhanced while the Macro-F1 scores are significantly degraded, which indicates that the accumulation of the classification errors is very high during this period. These findings clarify why the obtained Micro-F1 scores by CTWO method are declined in the following iterations.

Again, our proposed approach has conducted a better-balanced learning, although that the imbalance problem is also triggered by the decrease of the Macro-F1 scores beginning from iteration 20. This is clearly shown from iteration 20 to iteration 30 wherein the obtained Micro-F1 scores by our approach are more significantly improved, while its Macro-F1 scores are less decreased, as compared to CTWO method. These results demonstrate that, by exploiting synthetic minority samples, the quality of the selected confident samples is better, and the accumulation of the misclassification errors is smaller during this period. As a consequence, our approach yields to the peak Micro-F1 score at iteration 40, and keeps superior results during the last 40 iterations, despite the performance degradations resulting from cumulative misclassification errors.

Overall, we conclude that, under different labeling ratios, the integration of the oversampling technique deems useful for reducing the label imbalance problem, which is exaggerated with the progress of the Co-training learning. Moreover, although that our approach does not lead to much significant performance improvement at the end of the Co-training learning as compared to CTWO method, it guarantees a better-balanced learning of the different labels along the different iterations, which allows to sustain a beneficial Co-training learning for a longer period.

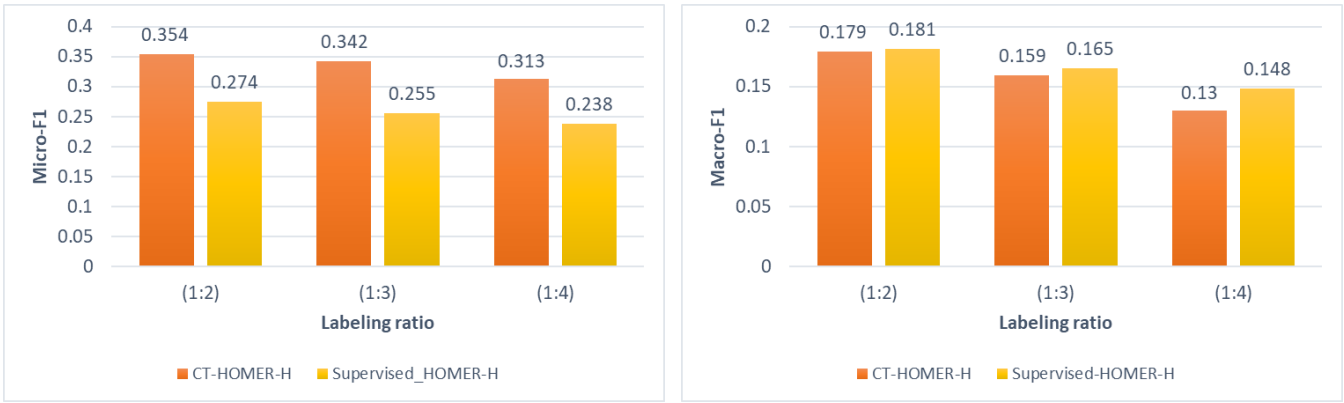
- **Does our suggested Co-training framework succeeds in improving the performance results when changing the base hierarchical multi-label learner?**

The goal of this section is to verify that the proposed SSL framework can serve for improving the performance scores using other kinds of local HMC learners. Here, we use a modified version of HOMER method (Tsoumakas et al. 2008), denoted to as HOMER-H, that functions on a predefined label hierarchy, which makes it a local HMC method. More precisely, HOMER-H applies an LCPN strategy at the training step, and uses a top-down induction mechanism for deducing the final labels of unseen samples at the test step. So, we report the obtained results of the following supervised and semi-supervised baseline methods that use HOMER-H as a basic HMC learner:

- **Supervised-HOMER-H:** This method uses two HOMER-H classifiers, which are learnt using the initial labeled data (after rebalancing) while exploiting different data' views, and their outputs are combined based on the consensus principle, as in our proposed approach.

- **CT-HOMER-H:** a variant of the proposed Co-training-based approach, using HOMER-H classifier as a basic hierarchical multi-label learner.

In this experiment, each multi-label classifier used in HOMER-H is built using the well-known Binary Relevance (BR) method, that trains a single binary classifier for each label in the hierarchy. In both comparative methods, we keep using Random Forest as a basic binary classifier. Note that the implementation of the default HOMER and Binary Relevance (BR) methods is already provided in Mulan package.



(a) Micro-F1 scores

(b) Macro-F1 scores

FIGURE 4 Classification performances of Supervised-HOMER-H and CT-HOMER-H methods in terms of both evaluation measures at different sampling rates. The horizontal axis represents the different labeling ratios while the vertical axis indicates the used evaluation measure.

Figure 4 presents the obtained results of both methods (i.e. Supervised-HOMER-H and CT-HOMER-H) in terms of both evaluation measures, and at different labeling rates. Based on this figure, several conclusions can be made. First, the semi-supervised CT-HOMER-H method usually leads to better performances in terms of Micro-F1 scores as compared to Supervised-HOMER-H method, which uses the labeled data only. These results indicate that our suggested Co-training framework can be successful in improving the overall predictive capability when using other kinds of HMC learners. Nonetheless, in terms of Macro-F1 scores, CT-HOMER-H method gets lower performances as compared to those obtained by Supervised-HOMER-H method. These results indicate that CT-HOMER-H also suffers from the exaggeration of the label imbalance issue during the Co-training learning, which yields to the decrease of the Macro-F1 scores.

Second, by looking at Figure 2, we deduce that our Co-training-based approach is constantly outperformed by CT-HOMER-H method in terms of Micro-F1 and Macro-F1 scores at different labeling ratios. Moreover, we observe that Supervised-HOMER-H method usually leads to higher performances than the supervised baseline, which uses the adopted hierarchical multi-label classifier in our approach (described in section 3.2), in terms of both evaluation measures. Obtained results reveal that HOMER-H learner performs better than the used hierarchical multi-label classifier in our proposed approach.

Finally, we can deduce that higher performances are expected to be achieved by our approach using more efficient HMC learners with higher classification performance, and better capability in handling the label imbalance issue.

5 | CONCLUSIONS

In this paper, we propose a semi-supervised approach, which is based on the Co-training paradigm, for the HMC task of scientific papers to efficiently tackle the scarcity issue of labeled training data. Our Co-training-based approach exploits content and bibliographic coupling information as two distinct views of scientific papers. Based on these two views, two hierarchical multi-label classifiers are iteratively trained to teach each other with their most confident predictions on unlabeled samples. To avoid adding noisy unlabeled samples during the iterative learning process, we use two suggested criteria for confident samples selection, which are: Maximum Agreement and Labels Cardinality Consistency. Moreover, our approach applies an oversampling technique to rebalance the labels distribution in the initial labeled set by the addition of synthetic minority samples. Performed experiments validate the efficiency of the proposed approach in improving the performance scores over the supervised baseline. Our experiments have also shown that MLSMOTE guarantees a better-balanced learning that leads to better final performance scores. However, our approach is still suffering from performance degradation at the latest Co-training rounds.

As future work, we plan to study whether the incorporation of other imbalance-aware techniques (e.g. feature selection, cost-sensitive learning), can improve the performance results of our suggested Co-training-based framework.

References

Aljaber, B., Martinez, D., Stokes, N., & Bailey, J. (2011). Improving mesh classification of biomedical articles using citation contexts. *Journal of biomedical informatics*, *44*(5), 881–896.

Baker, S., & Korhonen, A.-L. (2017). Initializing neural networks for hierarchical multi-label text classification..

Blockeel, H., Bruynooghe, M., Džeroski, S., Ramon, J., & Struyf, J. (2002). Hierarchical multi-classification. In *Workshop notes of the kdd'02 workshop on multi-relational data mining* (pp. 21–35).

Blockeel, H., Schietgat, L., Struyf, J., Džeroski, S., & Clare, A. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics. In *European conference on principles of data mining and knowledge discovery* (pp. 18–29).

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 92–100).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Cao, M. D., & Gao, X. (2005). Combining contents and citations for scientific document classification. In *Australasian joint conference on artificial intelligence* (pp. 143–152).

Caragea, C., Bulgarov, F., & Mihalcea, R. (2015). Co-training for topic classification of scholarly data. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2357–2366).

Cerri, R., Barros, R. C., & de Carvalho, A. C. (2012). A genetic algorithm for hierarchical multi-label classification. In *Proceedings of the 27th annual acm symposium on applied computing* (pp. 250–255).

Cerri, R., Barros, R. C., & De Carvalho, A. C. (2014). Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, *80*(1), 39–56.

Cerri, R., Basgalupp, M. P., Barros, R. C., & de Carvalho, A. C. (2019). Inducing hierarchical multi-label classification rules with genetic algorithms. *Applied Soft Computing*, *77*, 584–604.

Chapelle, O., & Zien, A. (2006). *B, scholkopf. semi-supervised learning*. MIT press.

Charte, F., Rivera, A., del Jesus, M. J., & Herrera, F. (2013). A first approach to deal with imbalance in multi-label datasets. In *International conference on hybrid artificial intelligence systems* (pp. 150–160).

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397.

Chen, B., & Hu, J. (2012). Hierarchical multi-label classification based on over-sampling and hierarchy constraint for gene function prediction. *IEEJ Transactions on Electrical and Electronic Engineering*, *7*(2), 183–189.

Clare, A., & King, R. D. (2003). Predicting gene function in saccharomyces cerevisiae. *Bioinformatics*, *19*(suppl_2), ii42–ii49.

Costa, E. P., Lorena, A. C., Carvalho, A. C., Freitas, A. A., & Holden, N. (2007). Comparing several approaches for hierarchical classification of proteins with decision trees. In *Brazilian symposium on bioinformatics* (pp. 126–137).

Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Ziviani, N., Moura, E., & Ribeiro-Neto, B. (2006). A comparative study of citations and links in document classification. In *Proceedings of the 6th acm/ieee-cs joint conference on digital libraries* (pp. 75–84).

Couto, T., Ziviani, N., Calado, P., Cristo, M., Gonçalves, M., de Moura, E. S., & Brandão, W. (2010). Classifying documents with link-based bibliometric measures. *Information Retrieval*, *13*(4), 315–345.

Feng, S., Fu, P., & Zheng, W. (2018). A hierarchical multi-label classification method based on neural networks for gene function prediction. *Biotechnology & Biotechnological Equipment*, *32*(6), 1613–1621.

Irsan, I. C., & Khodra, M. L. (2019). Hierarchical multi-label news article classification with distributed semantic model based features. *International Journal of Advances in Intelligent Informatics*, *5*(1), 40–47.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, *3*(1), 1793–8201.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10–25.

Laguna, V. A., & de Andrade Lopes, A. (2009). A multi-view approach for semi-supervised scientific paper classification. *WAAMD*, *9*, 26–33.

Li, X., & Guo, Y. (2013). Active learning with multi-label svm classification. In *Twenty-third international joint conference on artificial intelligence*.

Liu, R.-L. (2017). A new bibliographic coupling measure with descriptive capability. *Scientometrics*, *110*(2), 915–935.

Łukasik, M., Kuśmierczyk, T., Bolikowski, Ł., & Nguyen, H. S. (2013). Hierarchical, multi-label classification of scholarly publications: modifications of ml-knn algorithm. In *Intelligent tools for building a scientific information platform* (pp. 343–363). Springer.

Metz, J., Freitas, A. A., et al. (2009). Extending hierarchical classification with semi-supervised learning. In *Proceedings of the uk workshop on computational intelligence* (pp. 1–6).

Otero, F. E., Freitas, A. A., & Johnson, C. G. (2010). A hierarchical multi-label classification ant colony algorithm for protein function prediction. *Memetic Computing*, *2*(3), 165–181.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Porter, M. F. (2006). An algorithm for suffix stripping. *Program*.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Roul, R. K., & Sahoo, J. K. (2017). Classification of research articles hierarchically: a new technique. In *Computational intelligence in data mining* (pp. 347–361). Springer.

Rousu, J., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2006). Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, *7*(Jul), 1601–1626.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Santos, A., & Canuto, A. (2014). Applying semi-supervised learning in hierarchical multi-label classification. *Expert systems with applications*, *41*(14), 6075–6085.

Santos, A. P., & Rodrigues, F. (2009). Multi-label hierarchical text classification using the acm taxonomy. In *14th portuguese conference on artificial intelligence (epia)* (pp. 553–564).

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1-2), 31–72.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, *24*(4), 265–269.

Sousa, R. T., & Gama, J. (2017). Comparison between co-training and self-training for single-target regression in data streams using amrules.

Sun, L., Kudo, M., & Kimura, K. (2016). A scalable clustering-based local multi-label classification method. In *Proceedings of the twenty-second european conference on artificial intelligence* (pp. 261–268).

Surkis, A., Hogle, J. A., DiazGranados, D., Hunt, J. D., Mazmanian, P. E., Connors, E., … others (2016). Classifying publications from the clinical and translational science award program along the translational research spectrum: a machine learning approach. *Journal of translational medicine*, *14*(1), 235.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, *3*(3), 1–13.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2008). Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ecml/pkdd 2008 workshop on mining multidimensional data (mmdâĂŹ08)* (Vol. 21, pp. 53–59).

Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning* (pp. 406–417).

Valentini, G. (2010). True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(3), 832–847.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine learning*, *73*(2), 185.

Vogrinčič, S., & Bosnić, Z. (2011). Ontology-based multi-label classification of economic articles. *Computer Science and Information Systems*, *8*(1), 101–119.

Wang, Q., Li, W., & Gool, L. V. (2019). Semi-supervised learning by augmented distribution alignment. In *Proceedings of the ieee international conference on computer vision* (pp. 1466–1475).

Wang, S., Minku, L. L., & Yao, X. (2018). A systematic study of online class imbalance learning with concept drift. *IEEE transactions on neural networks and learning systems*, *29*(10), 4802–4821.

Wei, Z., Wang, H., & Zhao, R. (2013). Semi-supervised multi-label image classification based on nearest neighbor editing. *Neurocomputing*, *119*, 462–468.

Wu, X.-Z., & Zhou, Z.-H. (2017). A unified view of multi-label performance measures. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 3780–3788).

Xing, Y., Yu, G., Domeniconi, C., Wang, J., & Zhang, Z. (2018). Multi-label co-training. In *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 2882–2888).

Zhan, W., & Zhang, M.-L. (2017). Inductive semi-supervised multi-label learning with co-training. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1305–1314).

Zhang, B., Chen, Y., Fan, W., Fox, E. A., Gonçalves, M. A., Cristo, M., & Calado, P. (2005). Intelligent fusion of structural and citation-based evidence for text classification. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval*

(pp. 667–668).

Zhang, M.-L., & Zhou, Z.-H. (2011). Cotrade: confident co-training with data editing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *41*(6), 1612–1626.

Zhao, Y., Baldini, I., Sattigeri, P., Padhi, I., Lee, Y. K., & Smith, E. (2018). Data driven techniques for organizing scientific articles relevant to biomimicry. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 347–353).