A systematic review of label correlations in multilabel learning

ANDRÉ DA SILVA PRADO, University of São Paulo, Brazil
ARIANE MACHADO LIMA, University of São Paulo, Brazil

CCS Concepts: • **Computing methodologies** → **Machine learning**;

Additional Key Words and Phrases: multilabel, label correlation

ACM Reference Format:
André da Silva Prado and Ariane Machado Lima. 2022. A systematic review of label correlations in multilabel learning. *J. ACM* 37, 4, Article 111 (August 2022), 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Multilabel problems are ubiquitous in real-world applications where multiple labels can be assigned to one instance. As a result, a plenty of algorithms have been proposed. Multilabel learning (MLL) can be seen as a generalization of traditional multiclass learning (MCL). While in MCL the labels are mutually exclusive since an instance can only belong to one class, in MLL the instances may simultaneously belong to more than one class [47]. Learning from multilabeled instances is to find a mapping function from the space of features to the space of labels.

Since instances can be associated with various labels, such labels may have correlations among them. Exploring label correlations (LC) is essential for improving prediction, because they are an important source of information. How to effectively exploit the underlying LC is a crucial task for MLL.

LC can occur in different forms, intensities and applications. For instance, a song can be classified "blues" and "rock" at the same time, as well as a movie can be labeled as "action", "adventure" and "science fiction". In image annotation, if the label "beach" is present in the image, it is very likely that the label "sea" will also show up, but unlikely this image will receive the label "tiger".

An important aspect of MLL lies in the fact that the number of possible label combinations grows exponentially with the number of labels [30], impacting efficiency and scalability of the learning algorithms when the number of labels becomes large [61]. Therefore, MLL methods that work in the original label space can easily become computationally impractical. Considering that there is redundant information among labels and the they are globally correlated with each other, LC can be used in methods for dimensionality reduction in label space, expecting to improve the classification accuracy and to reduce the training efforts [68].

Multilabel learning has attracted the interest of researchers. As a consequence, many reviews in MLL have being done. Some reviews mainly focused on a general overview of MLL algorithms [13, 43, 78, 102]. There are also reviews focused on MLL applied to specific problems, such as text

Authors' addresses: André da Silva Prado, prado.andre@usp.br, University of São Paulo, São Paulo, São Paulo, Brazil; Ariane Machado Lima, University of São Paulo, São Paulo, São Paulo, Brazil, ariane.machado@usp.br.

0004-5411/2022/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

topic models [2, 41], image annotation [28] and data stream classification [105]. Some reviews delved into specific multilabel algorithm families, such as genetic algorithms [14] and ensembles [48]. Some challenges of MLL were surveyed by Siblini et al. [58], whose review focused on dimensionality reduction, by Tarekegn et al. [63], whose focus were on imbalance and by Nieuwenhuis et al. [51], that addressed the optimization of MLL algorithms. Charte [4] surveyed the available MLL software tools. To the best of our knowledge, there is no review focused on MLL algorithms exploiting the label correlations.

The objective of this systematic review is to identify the state of art of multilabel methods that consider label correlation, in order to answer the following questions:

- what methods, performance measures and databases are being used for multilabel learning considering label correlations?
- what types of label correlations are being considered and how they are measured?

In addition, this review addresses two challenges in multilabel learning: 1) how to learn a multilabel model and 2) how to obtain a labeled training sample large enough to represent the space of all label combinations. Most of the proposed solutions for the first challenge are based on data transformation or method adaptation, in which label correlations have been used to improve the classifier performance. Solutions for the second challenge are based on improvements and augmentation of training samples using semi-supervised techniques that explore LC to recover the incomplete labels. Approaches to deal with these two challenges are presented in sections 5 and 6, respectively.

Figure 1 depicts an overview of this article that, in addition to the introduction, is organized as follows. Section 2 describes the methods used to perform this review. Section 3 discusses the three different aspects that the label correlations can be handled. Section 4 describes some approaches used to deal with label correlations that are worthy to be highlighted, because they are used in many of the selected articles. Section 5 explores the learning approaches for multilabel classification. Section 6 describes the semi-supervised strategies to improve the training dataset in order to increase performance. Section 7 presents a description of the most used datasets. Section 8 lists all used metrics and a brief discussion of the three most used. Finally, sections 9 and 10 present a global discussion and final conclusions.

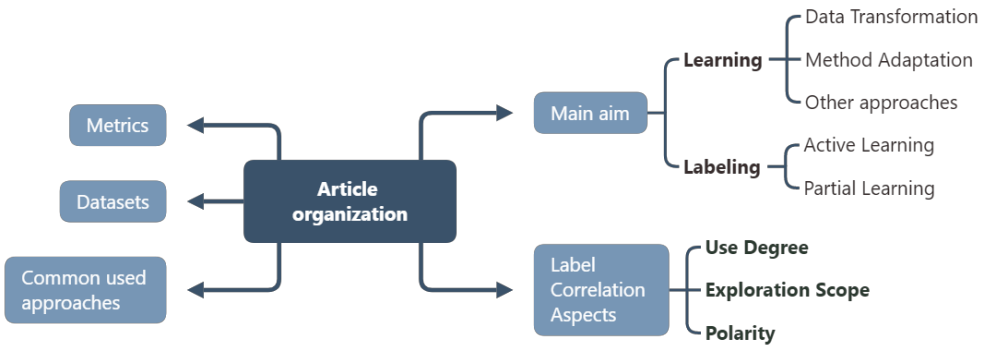

Fig. 1. Article overview

2 METHODS

A search string based on the words "multilabel", "label correlation", "classification" and their respective synonyms was submitted to the scientific repositories ACM Digital Library[1], IEEE Xplore[2], Scopus[3] and Web of Science[4] at March 9th, 2022. In order to focus on the most recent studies, the searches were performed considering articles published since 2016. The specific queries used in each repository are available in the appendix A.

Figure 2 summarizes the selection process. From the 782 articles returned from the repositories, after removing duplicates, inclusion and exclusion criteria were applied based on article's title and abstract.

The inclusion criterion was:

(1) articles that propose new multilabel learning methods that explore the label correlations.

The exclusion criteria were:

(1) articles that did not fully describe the used technique or algorithm;
(2) articles not fully available on the internet;
(3) articles that only used algorithms proposed in previous articles;
(4) articles not written in English;
(5) articles that did not mention the used datasets;
(6) articles that only addressed improvements in identifying clusters of labels;
(7) articles whose main problem was multi-instance learning;
(8) articles whose main approach was multi-view or multimodal learning;
(9) articles that proposed only feature selection methods for multilabel datasets.

Finally, on the 246 articles that adhere to the inclusion but to no one exclusion criterion, quality criteria were applied in order to select the most relevant articles in terms, for instance, of reproducibility. After the whole process, only articles that matched all the four quality criteria were selected, totalizing 92 articles that were included in this review.

The quality criteria were:

(1) presence of performance evaluation using publicly available datasets;
(2) presence of flowcharts or other visual support in the method presentation;
(3) presence of the pseudo-code of the algorithm;
(4) presence of tables, graphs or other visual support in the results presentation.

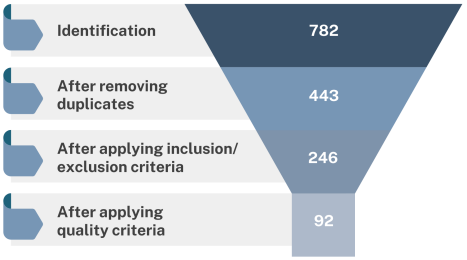

Fig. 2. Selection process

[1] dl.acm.org, searches at February 18th, 2021 and at March 9th, 2022
[2] ieeexplore.ieee.org, searches at February 17th, 2021 and at March 9th, 2022
[3] scopus.com, searches at February 18th, 2021 and at March 9th, 2022
[4] webofscience.com, searches at February 18th, 2021 and at March 9th, 2022

3 LABEL CORRELATIONS ASPECTS

LC can be analyzed by different aspects depending, as shown in figure 3 and discussed in this section. The correlation degree of use refers to the number of labels considered for correlation computation: whether the correlations are calculated in pairs of labels, among groups of labels or if there is no correlation (or it is not taken into account). The exploration scope analyses if label correlations are homogeneous in the entire dataset or change in different subsets. Finally, the correlation polarity distinguishes positive relations, where the occurrence of one label augment the occurrence of others, from negative relations, where the opposite is observed.

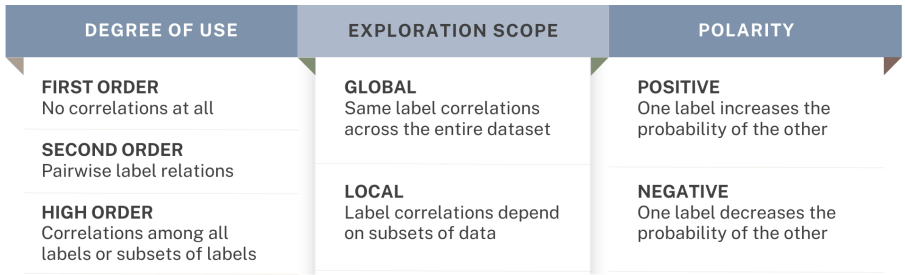

Fig. 3. Label correlations aspects

Figure 4 shows the proportion of the included articles that intentional and explicitly exploited each label correlation aspect, or only implicitly considering LC in a more generic way.

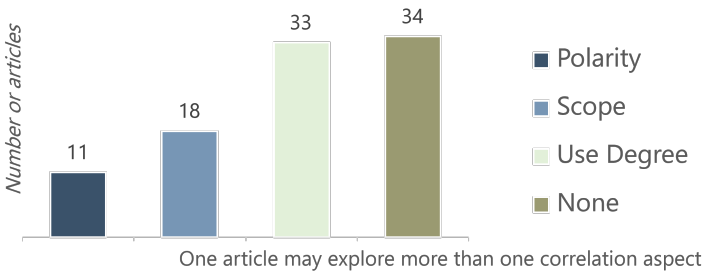

Fig. 4. Explicit exploration of correlation aspects

3.1 Label correlation degree: first, second or high-order

The LC degree in multilabel learning can be classified into three groups, from completely ignoring the existence of label correlation to considering that all labels can be correlated with each other.

3.1.1 First-order label correlation. First-order strategies consider that labels are independent, ignoring label correlations. This strategy tackles the multilabel classification problem decomposing it into a set of independent binary classification problems. It is simple to implement, but it cannot achieve satisfactory performance when data has indeed label correlation. As long as this review aims to analyze articles that explore LC, articles that only use first-order label relation were not included.

3.1.2 Second-order label correlation. Second-order strategy exploit pairwise label correlations, mostly by mining correlations between any pair of labels [20, 25, 29, 49, 88, 94, 103]. Lou et al. [44] designed a mechanism based on fuzzy inference to measure the correlation between two labels. Li et al. [33] employed a fully-connected and pairwise label graph.

When compared to first-order strategies, the efficiency and scalability of second-order methods suffer from the quadratic computational complexity with respect to the number of labels when label space becomes large [61].

3.1.3 High-Order. Real-world applications often present more complex relationships among the labels. In these cases, second-order methods may be ineffective [93]. High-order methods consider that the correlation may exist among all labels or label subsets. These are high processing demanding methods, however they can fully exploit the correlations existent in the data.

A common practice is to project the original label space to a low-dimensional label space. For example, if an instance is associated with multiple labels, there must be some sub-spaces of the instance space that are shared by these labels [61].

Some multilabel methods, such as classifier chain, stacking and label powerset (described in section 5.1), implicitly account for high-order correlations, once they consider more than pairs of labels during the learning process.

As examples of others approaches that can capture high-order LC, Shi et al. [56] and Shi et al. [57] used a neural network, Wu et al. [80] employed low-rank constraints and Gweon et al. [19] used the probability outputs of independent classifiers as input in the label space, Wu et al. [87] explored high-order LC by building a hierarchical tree structure, and Sun et al. [61] employed hypernetwork, where hyperedges are viewed as subspaces of instance space where the LC are exploited.

3.2 Exploration scope: global or local

The exploration scope means what set of data will be considered to calculate the level of LC. If LC is calculated considering the whole dataset, this is the global exploration, also called unconditional or asymmetric label dependency. However, if the LC exploration is restricted to some subsets, the exploration scope is local, likewise termed as conditional or symmetric label dependency.

3.2.1 Global. Zhang et al. [100] exploited the global relation via a sparse representation method to complement the missing labels in a partial learning problem and used similarity among labels to exploit local relations. Sun et al. [62] captured the global relation with a coefficient matrix inspired in low-rank representation.

Although many algorithms mainly exploit LC globally [77, 98, 103], exploiting only global LC may lead to unnecessary and error predictions [23]. Therefore, it is common that authors exploit local relation, as described in the next section.

3.2.2 Local. Sun et al. [62] and Guan and Li [15] explored the local relations by using a label manifold regularizer. Chen et al. [7] considered the labels that appear in an instance as correlated and all other labels as uncorrelated. Wu et al. [85] used conditional label dependence for partial multilabel learning. Nazmi et al. [50] decomposed the training data into multiple disjoint clusters to build a distinct label similarity graph for each cluster, that is able to capture local LC.

3.3 Correlation polarity: positive or negative

The LC polarity considers the conditional probability of a label given a label already shown. If the probability increases, the relation is positive, otherwise it is negative. Therefore, two or more

labels are positively related when the presence of a label in an instance augment the probability of showing other specific labels.

It is intrinsic to exploit the positive LC in any second-order or high-order strategy, however labels are not only positively correlated, since some of them can be mutually exclusive. The negative correlation between two labels, for instance, means that if an instance is associated with one label, it is less likely to be associated to the other [23, 53]. Che et al. [5] used a nonlinear classifier based on random forests to account label dependence as well as a weighted classification loss separating false from negative positives. Wu et al. wrote three articles [82–84] using chi-square estimation to consider all possible positive and negative LC's. Huang et al. [23] explored positive and negative correlations for each training example by computing a maximum conditional probability. Nan et al. [49] used kNN to identify positive/negative LC. Rastogi and Mortaza [54] incorporated the negative LC by measuring the Euclidean distance between any two prediction vectors.

4 COMMON APPROACHES TO DEAL WITH LABEL CORRELATIONS

All of the algorithms that was listed in this systematic review deal with the correlations among labels. Among them, some techniques stand out because they were used by many authors in diverse multilabel approaches, so there are worth mentioning them.

4.1 Low rank and embedding

These techniques project the original space of features and/or labels into a latent space, where the hidden structure of the data can be identified and, by doing this, a compressed representation of the data is obtained. It can naturally capture LC, even in high dimensional data, and it is useful to deal with sparse data. Low rank and embedding are widely used for both labelling improvement of the training dataset and classification.

The low rank representation is one of the most used technique to exploit LC in MLL [11, 62, 80, 82, 89, 97] because it is able to create clusters of labels, identifying what label belongs to each label set, by grouping labels with dense similarities and separating in other groups labels whose resemblances are sparse.

It is also used for label compression, in which the cost function measures the fit between a given matrix (the data) and an approximating matrix (the optimization variable), subject to a constraint that the approximating matrix has reduced rank.

Frequently low rank is used in association with matrix factorization [59, 99]. Guo et al. [17] trained a low rank mapping matrix to exploit the relationship between the label space and the feature space. Wu et al. [81] used low rank to identify the noise subspaces from the noisy data, extract the noise and then calculate the noise level of each example-label pair. [65] and [64] selectively and jointly factorized the sample-label association matrices into products of individual and shared low-rank matrices and leverage this shared matrix to model LC.

The embedding layer operates in the latent space. Zhong et al. [106] modeled underlying LC leveraging a label embedding to adjust the penalties for correlated instances, enabling a better understanding of the data hierarchy in the label space, and alleviating the inference bias. Ma et al. [46], in order to exploit the relationships between features and labels, performed dimensionality reduction to find the informative shared features of input space and output space together, so the features are extracted in the latent low-dimensional subspace oriented by label information for mining the correlation between labels and features.

Some authors used dictionaries to improve the embedding techniques. Ding et al. [11] adopted an embedded semantic dictionary to encode features with recovered label matrix, that will be appropriately propagated to recover labels and improve LC. Jing et al. [27] exploited LC in input feature space and output label space, applying dictionary learning and a label consistency regularization

term in the input space and, in the output label space, a partial-identical label embedding, in which the samples with exactly same label set can cluster together and the samples with partial-identical label sets can collaboratively represent each other.

By grouping similar labels, the following articles can capture LC. Mei et al. [47] proposed to learn separated subspaces for features and labels by maximizing the independence between components in the label subspace to discover LC represented by independent label components. Chen et al. [7] built a label correlation embedding loss that encourages the relevant/positive label vectors to gather closely around the dummy cluster centroid and encouraging irrelevant/negative label vectors to locate far from the cluster centroid.

4.2 Pairwise label relation

The base of pairwise label relation is to calculate the quantity of instances that share the same label pairs.

Xu et al. [91] measured the level of correlation between a pair of labels using the cosine of label vectors. He et al. [20] proposed a partial pairwise label dependence, because label dependence is very sparse in some tasks, so exploiting fully pairwise label dependence might be unnecessary. Nan et al. [49] discovered the positive and negative pairwise LC from the training datasets, built a label powerset subclassifier and, in the test stage, identifyed the local LC of the unseen instances and employed the found LC to rectify the output.

4.3 Semantic correlation

When the label information is incomplete, the original label matrix cannot accurately reflect the relationship between the data features and the label categories. Huang and Zhao [26] used the semantic correlation between labels to select trainning data to the classifier and obtain a new complementary label matrix that contains more information. Zhang et al. [99] used the semantic-based and visual-based relations to recover label-to-label relation when there are cooccurrence of labels. Jing et al. [27] exploited the semantic correlation of the samples with exactly same label set and the samples with partial identical label sets.

4.4 Similarity measurements

Wang et al. [72] explored LC by enforcing distribution similarity between the predicted labels and the ground truth labels, minimizing the distance between the joint distribution of the predicted labels and the joint distribution of the ground truth labels. Koda et al. [30] added an extra term to penalize the dissimilarity among neighbors into the self-scaling variable metric objective function. Neighbors tend to share the same labels. Wang et al. [75] used the similarity to express the relationships between the instances more accurately through differentiating and integrating discrete and continuous features.

5 MULTILABEL LEARNING APPROACHES

The proposal of a multilabel learning approach was the focus of almost 70% of the articles analyzed in this systematic review (figure 5). Most of these approaches can be divided into two main categories: Data Transformation and Method Adaptation, presented in sections 5.1 and 5.2, respectively. The first transforms the data into multiple traditional binary classification problems, whereas the second adapts binary or multiclass algorithms in order to tackle the multilabel data. Some other proposals do not fit any of these categories, and are presented in section 5.3.

Figure 6 exhibits the number of articles proposing each multilabel learning approach per year. , year over year a greater number of published articles demonstrates that the interest of researchers in multilabel approaches that consider label correlations has been increasing.

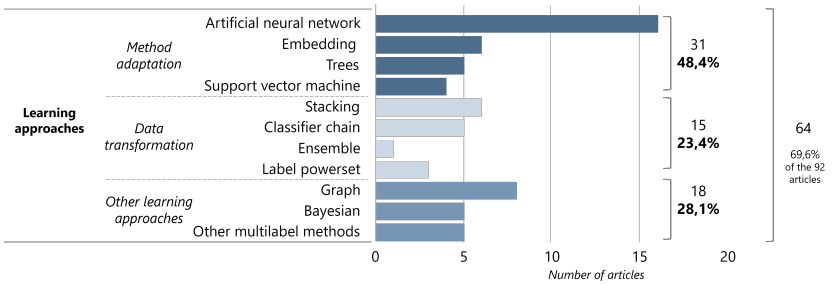

Fig. 5. Multilabel learning approaches

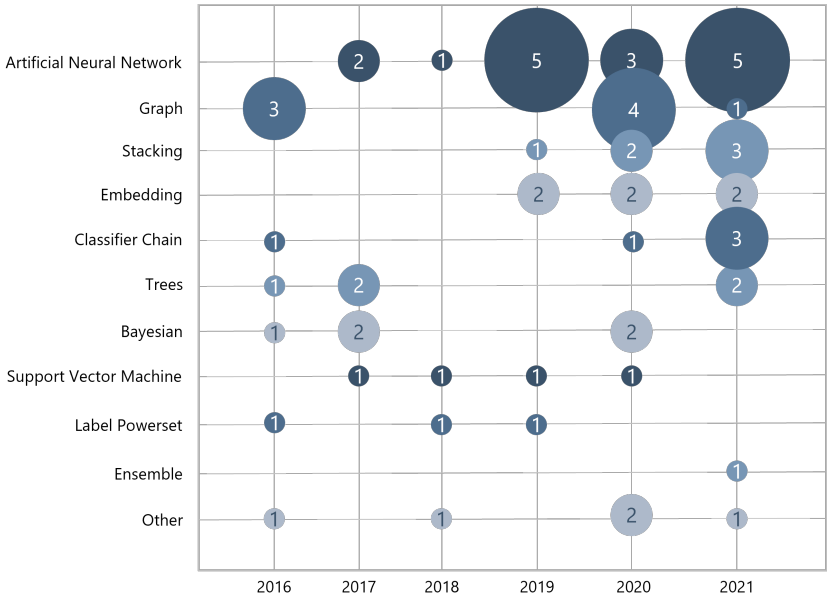

Fig. 6. Number of articles proposing each multilabel learning approach per year

5.1 Data transformation

Data transformation consists in decomposing a multilabel problem into several binary classification problems.

Binary Relevance (BR) is the simplest strategy for data transformation. This strategy transforms the multilabel problem into various independent binary sub-problems, where the presence or absence of each label is predicted by a binary classifier specifically trained for that label using the original data features. For a given instance, the predicted set of labels is the union of the results of all binary classifiers. BR has low computational cost that scales linearly with the number of labels and allows a simple parallel implementation. BR strategy does not consider label correlation, and so articles using purely this strategy are not present in this review. However it is important to remark that BR is the base method for most of data transformation algorithms, including those considering label correlation such as *classifier chain* and *stacking*.

5.1.1 Classifier chain (CC). This strategy applies a sequence of binary classifiers, where the input of each classifier receives the labels predicted by the previous classifiers as additional features. Formally, in a multilabel problem of L labels $l_1, ..., l_L$ for instances with F initial features $f_1, ..., f_F$, l_1 is predicted by a classifier that analyses instances $x^1 = (f_1, ..., f_F)$, and each one of the remaining labels l_i is predicted by a classifier that analyses instances $x^i = (f_1, ..., f_F, l_1, ..., l_{i-1})$, for $i = 2, ..., L$.

Such formulation allows the model to take into account label correlations, but only if the classifiers are chained in a suitable order. If the additional features $l_1, ..., l_{i-1}$ are highly correlated to the current label l_i, the performance will be improved, otherwise, the performance will not be influenced or even degraded [70]. How to identify LC and determine the label order is critical for the CC approach.

Wang et al. [69] bypassed the label order problem by employing a binary classifier to predict all labels simultaneously and applying an iterative reasoning mechanism to use the inter-label information, where each instance of reasoning takes the previously predicted likelihoods for all labels as additional input. Weng et al. [79] optimized the label ordering according to LC and select specific features from original feature space and label space. Lee et al. [31] built a directed acyclic graph (DAG) to capture LC, so that highly correlated labels can be sequentially ordered in the chains obtained from the DAG. Wang et al. [71] employed Bayesian network to model LC using conditional entropy, so the labels are sorted according to the topological network structure.

5.1.2 Stacking. This strategy is a BR-based learning strategy that trains two layers of classifiers. The first layer contains a standard binary classifier for each label. The second layer also learns a binary classifier for each label but uses, potentially, the labels predicted by all first layer classifiers as additional features. This point is essential to differentiate CC from stacking. While in CC the prediction of every label is decided jointly by the original features and the prior labels in the chain, in stacking the prediction of every label depends on the original features and potentially all the labels, eliminating the problem of label ordering typical of CCs.

A critical issue in this approach is that, when the predicted labels of the first layer are not correlated with the target label of the second layer, these additional features may bring noise into the second layer and reduce the learning performance. To overcome this problem two main strategies have being applied.

One strategy is selecting only a subset of labels of the first layer that are highly correlated with the target label in the second layer. This highly correlated subset was captured by Chen et al. [6] using F-score, Liu et al. [40] applied Relief-F, He et al. [20] used data-driven conditional probability with binary cross-entropy loss function and Weng et al. [77] proposed a concept that deal with multiples objectives called Pareto Optimum.

The second strategy is applying weights to reduce the effects of noisy and irrelevant labels. For this, Rastin et al. [53] used a feature-weighted distance measure based on positive/negative and global/local LC, and Xia et al. [88] employed and stacked ensemble strategy, considering the pairwise LC for determining weights of the ensemble members.

5.1.3 Label powerset (LP). This strategy transforms the multilabel data into multiclass data, where each label combination, called *labelset*, corresponds to a class. Thus, a labelset of an instance is the set of all labels associated to that instance. Therefore, a multiclass classifier can be trained to predict such labelsets. In this strategy the label correlations are implicitly considered because each labelset are, potentially, positively correlated labels. The disadvantage of this method is known as *label dimensionality*: for L labels, the number of labelsets is 2^L in the worst[5] case, which increases the

[5]Only the label combinations observed in the training sample are considered as a class.

model complexity, impacting negatively in the result. In addition, a training dataset with enough data to represent all label combinations is needed.

Li et al. [34] developed a method that forms two layers of labelsets that is able to filter and reduce the number of labelsets. The lower layer is built with subsets of the upper layer and it aim is describing label correlations within each labelset of the upper layer, .

Gweon et al. [19] and Nan et al. [49], instead of directly reduce the number of labelsets, aim to reduce the performance loss.

5.1.4 Ensemble. In ensemble methods, multiple classifiers are jointly applied with the premise that the union of several simple classifiers perform better than a single complex one, so diversity is the key of the ensemble robustness.

The ensemble strategy was applied by Wang et al. [73], that integrated multiple attention-based classifiers with a learned weight vector to produce a multilabel classifier. The label space was encoded into a low-dimensional pseudo-label space, and then decoded to the predicted full label space by learning the underlying correlation between pseudo-label and real label representations.

5.2 Method adaptation

Method adaptation is the approach that adapts existing machine learning algorithms to be used for multilabel learning (MLL). This category includes artificial neural networks and their derivations, decision trees, SVM and probabilistic methods.

5.2.1 Artificial neural networks (ANN). This method is among the most used methods for multilabel learning because they natively support multilabel problems. Each output node belongs to some label and outputs a score for that label. Neural Network models can exploit LC in the penultimate layer of the network.

Sihao et al. [59] proposed to use a single-hidden layer feed-forward ANN Extreme Learning Machine based on Label Matrix Factorization to account for the associations among labels and guarantee the algorithm complexity. They decompose the label matrix into a latent label matrix, perform the classification in the latent space with extreme learning machine and maps the predicted labels back to the original space. Hong et al. [22] employed neural tensor network to explore the relations among the labels of neighbors, using the *maximum a posteriori* principle, and use the labels of neighbors as feature to predict the label of the query instance.

Several articles used deep learning architectures. Huang and Zhao [26] proposed a residual network-based algorithm, where a classifier is trained for each label, using the semantic correlation between the labels to select training data for the classifier, avoiding bring missing labels in the sample. Chen et al. [7] used a deep learning framework to disentangle, embed and rank the corresponding label cues to explicitly model the LC. Wang et al. [66] exploited the LC using the hidden layer information in deep networks building the deep belief network as a single-label classifier for each class. LC are captured because the feature space of each binary classifier is extended with the hidden layer representation associations of all previous deep belief networks, as in a classifier chain. Shen et al. [55] proposed to learn prototype to represent each label, and preserve similarity among images, labels, and prototypes, further applying label-correlation aware loss on predicted label space to discover label correlation. Liu et al. [39] employed an attention mechanism and label correlation to deal with a large number of class labels, using matrix factorization to reduce the label, which encodes the latent space and preserve label correlation of original labels with the decode matrix. Wang et al. [68] proposed a deep learning based framework to connect the feature and label spaces, preserving local structure of feature space by graph regularizations and exploring LC by hypergraphs.

Convolutional Neural Networks (CNN) is a popular method for image recognition also in multilabel applications, and has been widely used for graph learning and representation. Zhang et al. [99] exploited label correlations by utilizing the semantic cooccurrence information of labels, applying CNN to reduce semantic GAP between the image visual content and semantic concepts. Xue et al. [92] built an image feature map using convolutional neural networks and then apply a squeeze-and-excitation block, which is a structural unit able to model correlations between image channels. Li and Yang [36], using a Graph Convolutional Network, designed a mapping function to better learn the correlation in the label space, via a label correlation matrix based on their co-occurrence patterns to describe the label dependencies. Li et al. [37] applied graph convolutional networks to built a label relation graph by thresholding on cosine similarity computed from mutual embedding similarities.

An *autoencoder* is a special type of feed forward artificial neural network that learns efficient codings (encoding) by training the network to ignore the noise in the data. Hidden layers represent the code used to encode the input and a decoder tries to reconstruct from the reduced encoding a representation as close as possible to its original input. It is frequently used for dimesionality reduction.

Lian et al. [38] reconstruct input matrices under three assumptions: multilabel independence, multilabel dependence and partial multilabel dependence to incorporate this information into deep supervised autoencoder. Cheng et al. in their articles [9] and [10] proposed the usage of kernel extreme learning machine autoencoders with non-equilibrium labels completion to capture the correlations in the labels space and in the feature space, using entropy to measure the relationship between the unknown labels and the known labels weight.

5.2.2 Embedding methods. Label embedding methods are used to represent discrete variables as continuous vectors. They jointly extract the information of all labels, aiming to transform the original label space into an low dimensional underlying space, reducing the output space and the computational complexity. They can exploit the hidden structure of the original space and make full use of LC.

Many authors used neural networks to apply label and/or feature embedding, as Wang et al. [68] that introduced a deep cross-view embedding that is able to correlate the feature space and label space at the same time. Shi et al. [56] built a two-layer network structure, with a high-level label–label network to learn LC and a low-level node–node network to learn node interactions within the same latent embedding space. Yao et al. [94] embeded each label as a vector of the weight matrix, calculate the initial values of a label embedding matrix to incorporate the label co-occurrence matrix into neural networks, capturing and exploring LC in the penultimate layer of ANN models. Mei et al. [47] proposed learning features and labels subspaces separately for maximizing the independence between components in the label subspace and so, discover LC represented by independent label components

Zhong et al. [106] proposed a label-correlated classification method, represented by a label embedding layer operating in the latent label space.

Liu et al. [39] used matrix factorization to perform the algorithm and consider label correlation of the original label space in this process

Huang et al. [25] modeled label correlation by learning an embedding matrix from a pre-defined label correlation graph by graph embedding and constructed a multilabel classifier from the low-dimensional latent feature space to the label space, where the embedding matrix is utilized as the model coefficients.

Su et al. [60] applied a deep matrix factorization to analyzing complex hierarchical and structural data, imposing a low-rank constraint. The algorithm learn an inverse covariance matrix and exploit it to capture the co-occurrence patterns among the latent attributes and among the given labels.

5.2.3 Support vector machines (SVM). Koda et al. [30] explored LC by means of a structured SVM, integrating the output structure and spatial information simultaneously during the training. Zhang et al. [101] incorporated into the predictive function a label specific component and assume that exists a clustered relationship between labels represented by the label-specific parts in support vector machine. He et al. [21] proposed a SVM-based method that builds a label covariance matrix to capture and explore LC and learns the model parameters corresponding to each label and virtual label simultaneously.

Wu et al. [80] proposed to joint Ranking SVM with Binary Relevance as a way to make use of the advantages of both – minimized ranking loss by means of Ranking SVM and seize simplicity of BR - using low rank to exploit LC.

5.2.4 Decision trees. One of the greatest advantages of decision trees is their interpretability, that is an important aspect in some applications. Trees, in MLL, are also capable to build hierarchical structures of the labels, capturing LC.

Wu et al. [87] and Chandran and Panicker [3] exploited LC constructing a set of hierarchical trees, that are combined into as an ensemble to do the prediction, where the higher levels of the trees contain the labels with the highest probabilities. Prati et al. [52] used generalized fuzzy entropy, aggregated over all labels, to choose the best attribute for growing the tree. It also can generate leaves predicting partial label sets, which can incorporate in the model some aspects of LC.

Antonucci and Corani [1] computed whether or not a class label is optimal, returning a compact description of the set of optimal sequences of labels. LC were shaped with a tree topology.

Wei et al. [76] developed an algorithm taking label similarity in account to learn from partially observed labels and update the tree always when new labels emerge.

Yao et al. [96] proposed to built a multilabel cluster tree by iteratively invoking a clustering method, then using the majority labels of instances to annotate each node in the tree. Once the tree is constructed, a multilabel crotch ensemble classifier is built with a weighted voting scheme.

5.3 Other learning approaches

There are approaches to tackle the multilabel problem that are not entirely data transformation neither method adaptation. The probabilistic theoretical framework is strongly used by many authors. Likewise, graphs are frequently used to represent the dataset, its instances, labels and correlations.

Data transformation and method adaptation are the most used strategies to tackle multilabel problems, nonetheless it is possible to employ different strategies. In this review it was possible to identify approaches that are not entirely any of them, that introduce other techniques into MLL. This section begins describing five specific approaches and highlights the use of graphs and probabilistic models, that were applied in many articles.

Xu et al. [91] used an information-theoretic semi-supervised learning for solving linear and nonlinear multilabel problems and propose algorithms with a cost-sensitive objective function, making use of the cosine of label vectors to characterize the correlation level between a pair of labels.

Wang et al. [72] proposed an adversarial framework for MLL that is able to implicitly capture the joint distribution of multiple labels that contains all LC.

Lou et al. [44] introduced a fuzzy system as the basis model, by which the hidden relationship between labels and features can be learned and label correlation measure mechanism based on fuzzy inference is designed to learn the label correlations.

Xu et al. [90] proposed to directly constraining the local Rademacher complexity in an empirical risk minimization based algorithm, so that a tighter generalization error bound of the algorithm could be expected. This new constraint achieves better recovery of the low-rank structure of the predictor and exploits LC.

Xie et al. [89] presented a multilabel transfer annotation system based on the perceptron criterion and use the Green's function to capture LC.

5.3.1 Graphs. Graph-based approaches represent the whole dataset as a graph. In It is a common practice in the graph multilabel structure to consider the labels as nodes and the edges as LC. Such edges are weighted by Wang et al. [67] with the conditional probabilities between labels within the dataset. Zhang et al. [104] regard each variable (observed features and hidden labels) as a node and second-order LC as edges, called factor functions. Then, the learning is made through maximizing the joint probability of the factor functions. Yao et al. [95] developed a semi-supervised learning approach that explore the pairwise LC by building a subgraph to capture pairwise LC via linear combination of their co-occurrence similarity and kernel-based similarity.

Sun et al. [61] represented high-order correlations among labels in hyperedges. LC are explicitly represented by the label vectors and weight vectors of hyperedges, where hyperedges are viewed as sub-spaces of instance space and label correlations are exploited from these sub-spaces.

Li et al. [35] proposed a SVM-Graph learning approach to explicitly models LC by learning a label correlation graph that reflects the underlying topological structures among labels.

Li et al. [32] leveraged graph convolutional networks with an adaptive label correlation graph to model LC, applying a Label Graph module to learn LC with word embeddings, utilizing GCN to map this graph into label-dependent object classifiers.

Wang et al. [75] built a graph-based algorithm which adopts kNN and random walk algorithms, where the random walk was used to explore LC through the connectivity between vertices on a graph model and the algorithm creates node sets containing only their kNN training instances and not necessarily the entire training set.

Nazmi et al. [50] proposed a classifier system that leverages a structured representation for the labels through undirected graphs to utilize the label similarities when evolving rules.

5.3.2 Probabilistic. These approaches aim to make predictions based on the probabilistic representation of uncertainty. Probabilistic methods are particularly useful when there is few training data. Probabilistic models are able to represent conditional models, that can be used to represent LC, i.e. calculate the probability of a label given one or several other label, so it is a theoretical framework that provides a tool for calculating LC.

The Bayesian models are the most common methods for probabilistic approaches, being the base model for many other.

Kim et al. [29] proposed a Bayesian method that uses pairwise LC for determining the most probable label set for a given unseen instance. Li et al. [33] developed a Bayesian-based method that learned image-dependent label structures by considering conditional label correlations as linear weight functions of features.

Huang et al. [23] built a Bayesian model based on the k-NN algorithm, making prediction through maximizing the posterior probability, which is estimated on the label distribution, the local positive and negative pairwise label correlations of the k-nearest neighbors.

Che et al. [5] proposed a framework based on hidden Markov model, built with a nonlinear classifier to model correlations among labels and time-dependency.

6 IMPROVEMENT OF DATASET LABELING

A problem that arises in multi-label classification is the dimensionality of labels. In a dataset with n possible labels, then there are 2 to the power of 9 combinations of those labels. So, in order to have a good learning, it is needed to have a dataset labeled as best and biggest as possible.

In many real-world applications, labeling is manual and depends on an expert, being an expensive and time-consuming process, specially when the label space is huge. For this reason, frequently there are many instances in which labels are missing, partially or even wrongly annotated, what decreases the performance of the classifier. Considering this reality, 28 articles in this revision (30.4% of the total) focus on improving the labeling and the size of the training base before starting the learning process, being 21 applying partial learning and 7 applying active learning techniques.

Partial Learning adapts a learning process for scenarios where training bases have many instances that are partially labeled, unlabeled, or even mislabeled. Active Learning selects which unlabeled instances from the training base are most informative for the classifier and asks the expert to label only those instances, reducing the workload.

6.1 Partial multilabel learning (PML)

In Partial learning, the ground-truth label is hidden within the candidate label set. Partial multilabel learning, also called by some authors as Weak Label Learning, is the problem where relevant labels of training data are partially known and many relevant labels are missing, resulting in abundant training data associated with an empty label set. The valid labels, as well as the false positive labels, are hidden in the candidate label set. The training procedure can be misguided by the noisy labels [62].

A general PML framework work in two steps. The first step eliminates irrelevant labels, where LC are widely used to recover the ground-truth label matrix. With the refined label space obtained from the first step, the second induces the classifier.

Most approaches, in order to explore LC, used low rank assumption on the observed label matrix, decomposing it into two matrices, for describing the latent factors of instances and labels, respectively, as it was in [93] and [98]. The latter also tackled the problem of PML high computational cost by means of label compression, that is able to optimize the efficiency.

Ding et al. [11] built a semantic dictionary to encode features with low-rank coding and link the learned codes with label input to recover the missing values and improve LC. Also, the article of Ma et al. [46] was based in partial learning and dictionary learning. In the output label space, low rank and sparse properties are used to perform the missing label imputation and in the input space, the features are extracted in the latent low-dimensional subspace oriented by label information for mining the correlation between labels and features.

Ye et al. [97], additionally to incomplete label, tackled the problem of corrupted features by decomposing the acquired feature matrix into an ideal feature matrix and an outlier matrix and utilizing the visual information among instances with a graph Laplacian regularization.

Rastogi and Mortaza [54] considered the structural property of data and the pairwise label correlations to recover missing labels. We have consider label-specific data features in order to discriminate each class label and used label-specific features for training the classifier.

Liu et al. [42] treated missing labels as the three states, being +1 for positive, -1 for negative and 0 for missing labels and used a Long Short-Term Memory to complete the label-level enhancement and extract the label correlation.

Guo et al. [18] proposed a two-classifier strategy to separately tackle frequent and infrequent labels, maintaining high performance on frequent labels and adopted a label correlation network to explore the label correlation with flexible aggregators.

Cheng et al. [8] introduced a two-level autoencoder, in which the first level is an autoencoder block based on the kernel extreme learning machine, whose information of labels is added to the input layer, generating in the output layer features that contain the correlation between features and labels. The second layer is the classification module, which uses the label completion matrix obtained by previous learning as a new label space.

Manifold regularization is a semi-supervised framework that exploits the geometry of the marginal distribution and incorporates labeled and unlabeled data in a classifier, embedding them as regularization terms. These aspects make manifold regularization particularly useful in datasets with few labeled data, because few labeled data are needed, once it uses a large quantity of unlabeled data.

Feng et al. [12] decomposed the observed label matrix into two matrices, one of them describing the latent factors of labels. These latent factors and LC are mutually adapted via label manifold regularization to exploit LC in an explicit manner. Guan and Li [15] proposed a generative Bayesian Model to tackle the problem of the incomplete labels by exploiting LC and further jointly leverage manifold regularization and label confidence constraint. Guan et al. [16] built joint learning of the latent ground-truth confidences and the prediction model, applying manifold constraints and low-rank to capture local label correlations and neighboring structure of instances, so as to accurately estimate the latent ground-truth confidences. Zhang et al. [100] used a manifold regularized sparse model to exploit LC and consider the feature structure, in order to conduct the complement of missing labels.

In Wang et al. [74] proposed a label embedding based on a two-level label recovery mechanism for incomplete datasets, in which the label space is projected to the inherent space where the instance and label correlations are captured. In [65] and [64], Tu et al. used low-rank matrix to deal with the noise in the input space and reduce the impact of unreliable annotators. Jing et al. [27] improved the dataset training that have much redundancy and noise by incorporating the dictionary learning to represent the input space, applying a partial-identical label embedding in the label output space, in which the samples with exactly same label set can cluster together, and the samples with partial-identical label sets can collaboratively represent each other.

Ma et al. [45] created a missing label recovery method using kNN-sparse graphs based on instance-wise smoothness and the label-wise smoothness, that can extract the sparse label dependency structures shared among labels, integrating the label imputation and the classifier training in one task.

Huang et al. [24] searched to discover completely unobserved labels with a clustering based regularization term, exploiting pairwise label correlations with cosine similarity.

Zhang et al. [103] proposed an ensemble learning that creates multiple graphs using data points as anchors based on randomly selecting a subset of features, using the intrinsic correlation among multiple labels to define the label correlation consistency term.

6.2 Active learning

Sampling is the foundation of active learning. Those algorithms select, by iterations, the most informative example model from partial labeled or even unlabeled data and updates the training dataset. A human specialist labels the instance and then the dataset is updated and the classifier trained again. By selecting only the most important instances, active learning aims to considerably reduce the labeling effort and cost.

Guo et al. [17] trained a low rank mapping matrix to capture a mapping relation between feature space and label space and selects partial high-informativeness example-label pairs from unlabeled example pool to perform automatic predictions. Wu et al. [82] took advantage of active adaptative learning method to learn feature representation from noisy data as the input instead of the noisy

data, capturing LC by means of low rank representation. [83] developed an example sampling strategy, which considers the measures of example noise together with example label uncertainty and LC to select informative examples and related labels for annotation. Wu et al. [81] developed an active learning algorithm based in low rank assumption for mitigating the influence of noisy labels in the training set.

Wu et al. [86] and Wu et al. [85] applyed active learning to MLL, exploring LC by means of conditional label dependency.

Wu et al. [84] combined, after human annotation, classification prediction information, example spatial information and LC to annotate some selected example-label pairs.

7 DATASETS

In this section, the most used datasets for validation of multilabel algorithms and their features will be discussed.

Ordinary public classification datasets are not useful for researches in the field of multilabel learning due to their binary or multiclass nature. However, with the increasing number of researches in this field, public multilabel repositories have been made available.

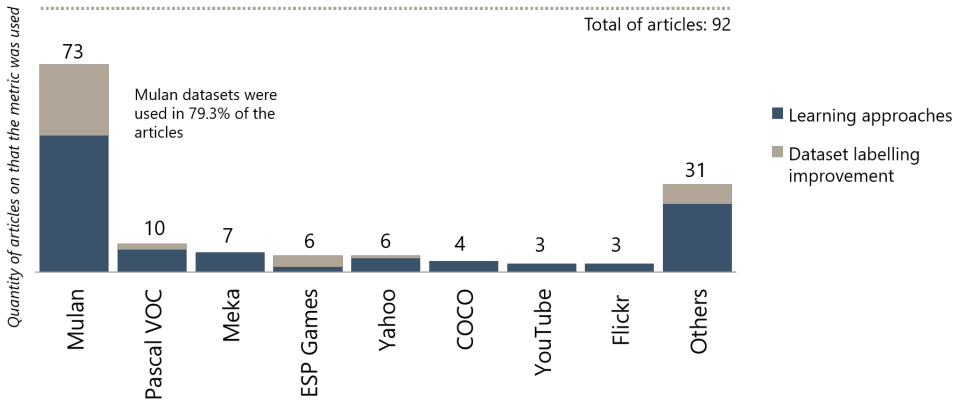

Fig. 7. Most used datasets

Mulan stands out as the most used dataset repository, once it was used in 79.3% of the articles of this review.

7.1 Mulan

Mulan is a Java Library for Multilabel Learning that includes 26 publicly available multilabel datasets from 5 domains, including audio, biology, image, text and video, some of them with thousands of labels, described in table 1. Due to these features, Mulan datasets are currently the most used datasets for multilabel learning.

In the download page[6] the attributes of each dataset are described in a clear way.

7.2 Meka

Meka is an open-source Java framework based on Weka library that provides tools for multilabel learning, including a collection of multilabel datasets, using the well documented Weka's ARFF (Attribute-Relation File Format). Train-test splits are also available for download.

[6]Mulan datasets: http://mulan.sourceforge.net/datasets-mlc.html

Table 1. Mulan datasets specifications

Name	Domain	Instances	Attributes		Labels	Cardinality	Density	Distinct
			nominal	numeric				
bibtex	text	7395	1836	0	159	2.402	0.015	2856
birds	audio	645	2	258	19	1.014	0.053	133
bookmarks	text	87856	2150	0	208	2.028	0.010	18716
CAL500	music	502	0	68	174	26.044	0.150	502
corel5k	images	5000	499	0	374	3.522	0.009	3175
corel16k (10 samples)	images	13811±87	500	0	161±9	2.867±0.033	0.018±0.001	4937±158
delicious	text (web)	16105	500	0	983	19.020	0.019	15806
emotions	music	593	0	72	6	1.869	0.311	27
enron	text	1702	1001	0	53	3.378	0.064	753
EUR-Lex (directory codes)	text	19348	0	5000	412	1.292	0.003	1615
EUR-Lex (subject matters)	text	19348	0	5000	201	2.213	0.011	2504
EUR-Lex (eurovoc descriptors)	text	19348	0	5000	3993	5.310	0.001	16467
flags	images (toy)	194	9	10	7	3.392	0.485	54
genbase	biology	662	1186	0	27	1.252	0.046	32
mediamill	video	43907	0	120	101	4.376	0.043	6555
medical	text	978	1449	0	45	1.245	0.028	94
NUS-WIDE	images	269648	0	128/500	81	1.869	0.023	18430
rcv1v2 (subset1)	text	6000	0	47236	101	2.880	0.029	1028
rcv1v2 (subset2)	text	6000	0	47236	101	2.634	0.026	954
rcv1v2 (subset3)	text	6000	0	47236	101	2.614	0.026	939
rcv1v2 (subset4)	text	6000	0	47229	101	2.484	0.025	816
rcv1v2 (subset5)	text	6000	0	47235	101	2.642	0.026	946
scene	image	2407	0	294	6	1.074	0.179	15
tmc2007	text	28596	49060	0	22	2.158	0.098	1341
yahoo	text	5423±1259	0	32786±7990	31±6	1.481±0.154	0.051±0.012	321±139
yeast	biology	2417	0	103	14	4.237	0.303	198

Table 2. Meka datasets

Dataset	L	N	LC	PU	Description and original source(s)
Enron	53	1702	3.39	0.442	A subset of the Enron Email Dataset, as labelled by the UC Berkeley Enron Email Analysis Project
Slashdot	22	3782	1.18	0.041	Article titles and partial blurbs mined from Slashdot.org
Language Log	75	1460	1.18	0.208	Articles posted on the Language Log
IMDB (Updated)	28	120919	2.00	0.037	Movie plot text summaries labelled with genres sourced from the Internet Movie Database interface, labeled with genres.

L - The number of predefined labels relevant to this dataset
N - The number of examples (training+testing) in the datasets
LC - Label Cardinality. Average number of labels assigned per document
PU - Percentage of documents with Unique label combinations

Meka datasets[7] are described in table 2. Different from MULAN, that indexes labels as the final attributes, MEKA indexes as the beginning.

7.3 PASCAL VOC

PASCAL is an acronym to Pattern Analysis, Statistical Modelling and Computational Learning. VOC stands for Visual Object Classes and refers to the project organized by ETHZ (Zurich Technology Federal Institute), University of Edinburgh, Microsoft Research Cambridge and University of Oxford,

[7]Meka datasets: https://waikato.github.io/meka/datasets/

in what image recognition annual challenges was launched, from 2005 to 2012. Currently there are no more challenges, however all datasets remain available for download[8].

All image datasets are grouped in Fully Annotated Databases, Partial Annotated Databases and Unannotated Databases, as described in table 3.

Table 3. Pascal VOC datasets

Category	Dataset name
Fully Annotated	TU Darmstadt (formerly ETHZ) UIUC Image for Car Detection VOC2005 Database: Dataset 1 VOC2005 Database: Testset 2 VOC2006 Database
Partially Annotated	Caltech MIT-CSAIL - Objects and Scenes
Unannotated	TU Graz-02 101 Object Categories

7.4 Yahoo

The Yahoo Webscope Program is an online library with the purpose to provide datasets for non-commercial use by academics and other scientists[9]. Table 4 describes Yahoo Webscope datasets that are directed to multilabel learning.

Table 4. Yahoo Webscope multilabel datasets

Dataset	Description
Yahoo Data Targeting User Modeling	A sample of user profiles and their interests at Yahoo webpages. Each user is represented as one feature vector and its associated labels. Each dimension of the feature vector quantifies a user activity with a certain interest category, calculated from user interactions with pages, ads, and search results, all of which are internally classified into these interest categories. The labels are derived in a similar way, based on user interactions with classified pages, ads, and search results during the test period. There exists a hierarchical structure among the labels, which is also provided in the data set.
Yahoo News Ranked Multilabel Corpus	This corpus provides the actual text so that the researchers can derive their own features that are good best for their algorithms. This corpus provides a ranking of labels for each document in terms of its importance.

7.5 COCO image dataset

COCO is an acronym for Common Objects in Context. Sponsored by Microsoft, Facebook, MightAI and Common Visual Data Foundation, COCO is a project that aims to advance the computer vision. It provides a large scale multilabeled image dataset[10] for object detection, segmentation and labeling, whose description can be seen in table 5

Table 5. COCO dataset

Images	Labeled images	Object instances	Object categories	Stuff categories	labels per image	People with keypoints
330K	200K	1.5 million	80	91	5	250K

[8]PASCAL VOC datasets: http://host.robots.ox.ac.uk/pascal/VOC/databases.html
[9]Yahoo Webscope: https://webscope.sandbox.yahoo.com/
[10]COCO dataset: https://cocodataset.org/

7.6 Other used datasets

Beyond these aforementioned, there are some other multilabel datasets used in the articles: YouTube, BlogCatalog, ESP Games, ActivityNet Captions, Meka, MSRC (provided with Microsoft Research in Cambridge), IMDB and DBLP, ImageNet, Reuters Corpus (RCV1-V2), AG-multi, 20newsgroup, Movie, LoveNakamura, LoveEkman, AppleNakamura, AppleEkman, Genbase, Fashion 550K, Brightkite, Gowalla, Tokyo@Foursquare and New York@Foursquare, Lamda and Arxiv Academic Paper Dataset (AAPD).

8 METRICS

Binary and multiclass learning consider that labels are mutually exclusive, so the label prediction for each instance can be correct or incorrect. In MLL various labels can be assigned for each instance, thus the binary concept of correct is not enough to evaluate the classification performance. For example, if the expected labels are A, B, D, E, F and the predicted labels are A, B, D and F, the prediction is partially correct. Having some labels correctly predicted and some wrong labels is preferred than a prediction with no correct labels. It is better to get four of the five labels rightly than predict all of them wrongly. Then, due too this MLL particularity, binary and multiclass metrics have to be adapted and also specific metrics are used.

Among them, three metrics stand out: hamming loss was used in 53%, macro F1 in 46% and micro F1 in 45% of the 92 articles analyzed in this review, as shown in figure 8.

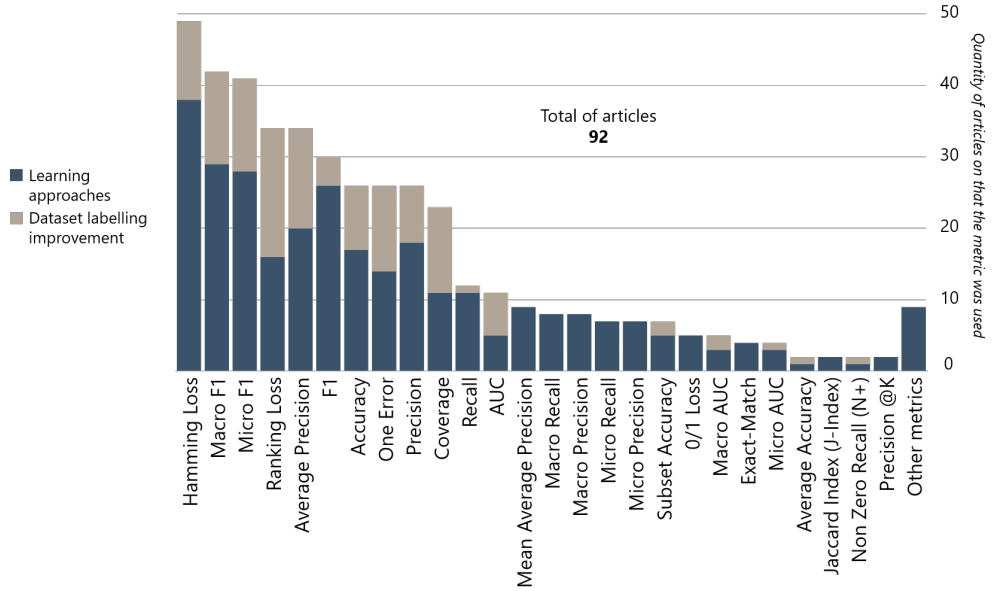

Fig. 8. Evaluation metrics for multilabel learning

For the following metric definitions, n is the number of instances; $L_i = L_i[1],...,L_i[l]$ stands for a set of class labels for the instance i; l is the number of labels; n is the number o instances; I is the indicator function; L_i^c are the predicted labels; L_i^d are the true labels labels. Taken from the matrix confusion, tp are true positives; tn are true negatives; fp are false positives and fn are false negatives. M denotes macro and μ denotes micro.

Hamming loss (equation 1) is the most used metric for multilabel learning, as described in the systematic review of label correlations in multilabel learning, in section 8. This loss metric takes into account the prediction errors, when a label is incorrectly predicted, and also missing errors, when a label is not predicted. The result is normalized over total number of classes and total number of examples. As it is a loss measurement, the lower the better.

$$HammingLoss = \frac{\sum_{i=1}^{n} \sum_{j=1}^{l} I\left(L_i^c[j] \neq L_i^d[j]\right)}{nl} \tag{1}$$

Some multilabel metrics are based on binary learning metrics, so a brief description of these binary metrics is needed.

Accuracy (equation 2) measures the proportion of correctly predicted values over the total predicted instances.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

Precision (equation 3) measures the proportion of true positives over the total number of positives predicted.

$$Precision = \frac{tp}{tp + fp} \tag{3}$$

Recall (equation 4) measures the proportion of true positives over the total positives predicted.

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

F1score (equation 5 is the harmonic mean between precision and recall.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

Many multilabel learning metrics are a generalization of binary learning metrics, to consider more than two classes averaging the metrics in some way.

Macro averaging ($Precision_M$, $Recall_M$,$F1score_M$) is an arithmetic mean of the *per class* metric over all classes, whose equations are 6, 7 and 8 respectively.

For each C_i class, there are true positives tp_i, false positives fp_i, true negatives tn_i and false negatives fp_i.

$$Precision_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l} \tag{6}$$

$$Recall_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l} \tag{7}$$

$$Fscore_M = \frac{2 * Precision_M Recall_M}{Precision_M + Recall_M} \tag{8}$$

Macro measurements do not take into consideration the imbalance of labels because they do not consider the quantity of times that a label appeared in the dataset, once these measurements calculates the average of the results per labels.

Micro average ($Precision\mu$, $Recall\mu$,$F1score\mu$) is the average overall metric of the classifier, whose equations are 9, 10 and 11.

$$Precision\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fp_i)} \tag{9}$$

$$Recall\mu = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)} \tag{10}$$

$$Fscore\mu = \frac{2 * Precision\mu Recall\mu}{Precision\mu + Recall\mu} \tag{11}$$

The average accuracy (equation 12 is the average per class effectiveness of a classifier.

$$AverageAccurary = \frac{\sum_{i=1}^{l} \frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l} \tag{12}$$

Exact-Match ratio evaluates how many times the predicted labels are exactly the same of the correctly labels, defined in equation 13:

$$ExactMatchRatio = \frac{\sum_{i=1}^{n} I\left(L_i^d = L_i^c\right)}{n} \tag{13}$$

Performance metrics can be obtained by different methodologies. A frequently used methodology is cross-validation, called k-fold, which consists of dividing the total dataset into k mutually exclusive subsets of the same size (folds). One subset is taken for testing and the remaining k-1 are used for parameter estimation.

9 DISCUSSION

By definition, labels are not mutually exclusive in MLL, thus multiple labels can be jointly assigned to the same instance. Considering this scenario, a label can often appear together with one or more specific labels in the same data instance, which demonstrates that there is a correlation among them. As identified in this systematic review, these correlations can be explored according to three fundamental aspects, which are the degree of use, the scope of exploration and the polarity. It is not necessary to explicitly explore all correlation aspects in an article. Even exploring just one or two of them, always the three aspects will be explored, implicitly.

In this review, whose focus is the correlation between labels, the only aspect compulsorily explored by the articles is the degree of use, that must be second-order or high-order, since the first-order ignore that exists LC.

Another point identified in this review is that the exploration of one aspect does not imply the exploration of another. For example, a high-order strategy does not imply exploring global scope, or negative polarity. This happens because the articles do not explicitly explore all aspects of correlation.

In MLL, LC is used with two very clear and distinct objectives. The first is to improve the performance of the classifier. As the possible combinations of labels grow exponentially according to the number of labels, the use of LCs helps to identify the most relevant labels for each instance, by ranking the most relevant related labels or to contribute to the reduction of the dimensionality of the labels by prioritizing labels with positive polarity correlation and deprioritizing those with negative polarity.

Multilabel algorithms require a considerably larger training base than binary or multiclass ones due to the large number of labels and their combinations. However, in real world, manual labeling is time-consuming and an expensive task, resulting in a large number of unlabeled, partially labeled, or even incorrectly labeled instances. For this reason, a large number of articles aim to improve

the quality of the training dataset. In this sense, LC is used to identify missing labels in partially labeled or unlabeled instances. In semi-supervised approaches, LC assists in selecting the most informative instances to be sent to the specialist.

The label space has an intrinsic hierarchy that is explored for some ML strategies, sometimes captured in an explicit way, such as trees and graphs, sometimes identified in the hidden space, such as embedding and low rank. These hierarchies can be more explored by other algorithms, mainly when there is a great number of labels, in order to help the selection of subsets of labels in stacking, to define the classifiers order in classifier chain or to reduce the cardinality of labels in other approaches.

Considering the aspects of label correlation, the articles have been widely explored the degree of use second-order and high-order, and also the exploration scope global, with some exploring the local scope. However few articles explore explicitly the negative polarity of the LC, so this is one more way to further researches.

Combining the automated annotation with human annotation of labels has been applied by some researchers, however there is still opportunity to improve the automated annotation, because there are increasing sources of large volume of data, that are difficult to be annotated by humans.

10 CONCLUSION

This systematic review aimed to study articles whose focus was the label correlation in multilabel learning. It provided a comprehensive perspective of LC because we made a deep dive into the most recent published articles that explore LC in MLL, with the objective to understand what have been done to capture LC and how they have been used to enhance the classification performance and to improve the quality of training datasets. As a result, it was possible to identify the constituent aspects of the correlation between labels, the main applications of LC, which are the classification itself and the improvement of the training base through automatic labeling. It was also possible to identify some approaches that stand out among the multilabel algorithms. In addition, a survey was carried out on the main public multilabel datasets and the main metrics used.

REFERENCES

[1] A. Antonucci and G. Corani. 2017. The multilabel naive credal classifier. *International Journal of Approximate Reasoning* 83 (2017), 320–336. https://doi.org/10.1016/j.ijar.2016.10.006 cited By 3.

[2] Sophie Burkhardt and Stefan Kramer. 2019. A Survey of Multi-Label Topic Models. *SIGKDD Explor. Newsl.* 21, 2 (nov 2019), 61–79. https://doi.org/10.1145/3373464.3373474

[3] S. A. Chandran and J. R. Panicker. 2017. An efficient multi-label classification system using ensemble of classifiers. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. 1133–1136. https://doi.org/10.1109/ICICICT1.2017.8342729

[4] Francisco Charte. 2020. A Comprehensive and Didactic Review on Multilabel Learning Software Tools. *IEEE Access* 8 (2020), 50330–50354. https://doi.org/10.1109/ACCESS.2020.2979787

[5] Y. Che, Y. Zhu, and X. Shen. 2020. Multilabel Classification With Multivariate Time Series Predictors. *IEEE Transactions on Signal Processing* 68 (2020), 5696–5705. https://doi.org/10.1109/TSP.2020.3027277

[6] Y.-N. Chen, W. Weng, S.-X. Wu, B.-H. Chen, Y.-L. Fan, and J.-H. Liu. 2021. An efficient stacking model with label selection for multi-label classification. *Applied Intelligence* 51, 1 (2021), 308–325. https://doi.org/10.1007/s10489-020-01807-z cited By 0.

[7] Z. Chen, Q. Cui, X. Wei, X. Jin, and Y. Guo. 2020. Disentangling, Embedding and Ranking Label Cues for Multi-Label Image Recognition. *IEEE Transactions on Multimedia* (2020), 1–1. https://doi.org/10.1109/TMM.2020.3003779

[8] Y. Cheng, F. Song, and K. Qian. 2021. Missing multi-label learning with non-equilibrium based on two-level autoencoder. *Applied Intelligence* 51, 10 (2021), 6997–7015. https://doi.org/10.1007/s10489-020-02140-1 cited By 2.

[9] Y. Cheng, D. Zhao, Y. Wang, and G. Pei. 2019. Multi-label learning with kernel extreme learning machine autoencoder. *Knowledge-Based Systems* 178 (2019), 1–10. https://doi.org/10.1016/j.knosys.2019.04.002 cited By 10.

[10] Y.-S. Cheng, D.-W. Zhao, Y.-B. Wang, and G.-S. Pei. 2019. Multi-label Learning of Kernel Extreme Learning Machine with Non-Equilibrium Label Completion]. *Tien Tzu Hsueh Pao/Acta Electronica Sinica* 47, 3 (2019), 719–725. https:

//doi.org/10.3969/j.issn.0372-2112.2019.03.029 cited By 2.

[11] Z. Ding, M. Shao, S. Li, and Y. Fu. 2018. Generic Embedded Semantic Dictionary for Robust Multi-Label Classification. In *2018 IEEE International Conference on Big Knowledge (ICBK)*. 282–289. https://doi.org/10.1109/ICBK.2018.00045

[12] L. Feng, J. Huang, S. Shu, and B. An. 2020. Regularized Matrix Factorization for Multilabel Learning With Missing Labels. *IEEE Transactions on Cybernetics* (2020), 1–12. https://doi.org/10.1109/TCYB.2020.3016897

[13] Eva Gibaja and Sebastian Ventura. 2014. Multi-label learning: a review of the state of the art and ongoing research. *WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY* 4, 6 (NOV-DEC 2014), 411–444. https://doi.org/{10.1002/widm.1139}

[14] Eduardo Corrêa Gonçalves, Alex A. Freitas, and Alexandre Plastino. 2018. A Survey of Genetic Algorithms for Multi-Label Classification. In *2018 IEEE Congress on Evolutionary Computation (CEC)*. 1–8. https://doi.org/10.1109/CEC.2018.8477927

[15] Y. Guan and X. Li. 2020. Multilabel Text Classification With Incomplete Labels: A Safe Generative Model With Label Manifold Regularization and Confidence Constraint. *IEEE MultiMedia* 27, 4 (Oct 2020), 38–47. https://doi.org/10.1109/MMUL.2020.3022068

[16] Y. Guan, B. Zhang, W. Li, and Y. Wang. 2021. Semi-supervised partial multi-label classification with low-rank and manifold constraints. *Pattern Recognition Letters* 151 (2021), 112–119. https://doi.org/10.1016/j.patrec.2021.08.005 cited By 0.

[17] A. Guo, Jian Wu, V. S. Sheng, P. Zhao, and Z. Cui. 2017. Multi-label active learning with low-rank mapping for image classification. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. 259–264. https://doi.org/10.1109/ICME.2017.8019412

[18] Lan-Zhe Guo, Zhi Zhou, Jie-Jing Shao, Qi Zhang, Feng Kuang, Gao-Le Li, Zhang-Xun Liu, Guo-Bin Wu, Nan Ma, Qun Li, and Yu-Feng Li. 2021. Learning from Imbalanced and Incomplete Supervision with Its Application to Ride-Sharing Liability Judgment. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 487–495. https://doi.org/10.1145/3447548.3467305

[19] H. Gweon, M. Schonlau, and S.H. Steiner. 2019. Nearest labelset using double distances for multi-label classification. *PeerJ Computer Science* 5 (2019). https://doi.org/10.7717/PEERJ-CS.242 cited By 0.

[20] T. He, L. Zhang, J. Guo, and Z. Yi. 2020. Multilabel classification by exploiting data-driven pair-wise label dependence. *International Journal of Intelligent Systems* 35, 9 (2020), 1375–1396. https://doi.org/10.1002/int.22257 cited By 1.

[21] Z.-F. He, M. Yang, H.-D. Liu, and L. Wang. 2019. Calibrated Multi-label Classification with Label Correlations. *Neural Processing Letters* 50, 2 (2019), 1361–1380. https://doi.org/10.1007/s11063-018-9925-2 cited By 2.

[22] W. Hong, W. Xu, J. Qi, and Y. Weng. 2019. Neural Tensor Network for Multi- Label Classification. *IEEE Access* 7 (2019), 96936–96941. https://doi.org/10.1109/ACCESS.2019.2930206

[23] J. Huang, G. Li, S. Wang, Z. Xue, and Q. Huang. 2017. Multi-label classification by exploiting local positive and negative pairwise label correlation. *Neurocomputing* 257 (2017), 164–174. https://doi.org/10.1016/j.neucom.2016.12.073 cited By 30.

[24] Jun Huang, Linchuan Xu, Kun Qian, Jing Wang, and Kenji Yamanishi. 2021. Multi-label learning with missing and completely unobserved labels. *DATA MINING AND KNOWLEDGE DISCOVERY* 35, 3 (MAY 2021), 1061–1086. https://doi.org/{10.1007/s10618-021-00743-x}

[25] J. Huang, Q. Xu, X. Qu, Y. Lin, and X. Zheng. 2021. Improving multi-label learning by correlation embedding. *Applied Sciences (Switzerland)* 11, 24 (2021). https://doi.org/10.3390/app112412145 cited By 0.

[26] M. Huang and P. Zhao. 2021. Image multi-label learning algorithm based on label correlation. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. 606–609. https://doi.org/10.1109/ICCECE51280.2021.9342484

[27] X. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang. 2016. Multi-Label Dictionary Learning for Image Annotation. *IEEE Transactions on Image Processing* 25, 6 (June 2016), 2712–2725. https://doi.org/10.1109/TIP.2016.2549459

[28] A. Kalaivani and S. Chitrakal. 2013. Challenges amp; approaches in multi-label image annotation. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. 1–8. https://doi.org/10.1109/ICCCNT.2013.6726482

[29] H.-C. Kim, J.-H. Park, D.-W. Kim, and J. Lee. 2020. Multilabel naïve Bayes classification considering label dependence. *Pattern Recognition Letters* 136 (2020), 279–285. https://doi.org/10.1016/j.patrec.2020.06.021 cited By 1.

[30] S. Koda, A. Zeggada, F. Melgani, and R. Nishii. 2018. Spatial and Structured SVM for Multilabel Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* 56, 10 (2018), 5948–5960. https://doi.org/10.1109/TGRS.2018.2828862 cited By 12.

[31] J. Lee, H. Kim, N.-R. Kim, and J.-H. Lee. 2016. An approach for multi-label classification by directed acyclic graph with label correlation maximization. *Information Sciences* 351 (2016), 101–114. https://doi.org/10.1016/j.ins.2016.02.037 cited By 31.

[32] Q. Li, X. Peng, Y. Qiao, and Q. Peng. 2020. Learning label correlations for multi-label image recognition with graph networks. *Pattern Recognition Letters* 138 (2020), 378–384. https://doi.org/10.1016/j.patrec.2020.07.040 cited By 0.

[33] Q. Li, M. Qiao, W. Bian, and D. Tao. 2016. Conditional Graphical Lasso for Multi-label Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2977–2986. https://doi.org/10.1109/CVPR.2016.325

[34] X. Li, J. Ouyang, and X. Zhou. 2016. Labelset topic model for multi-label document classification. *Journal of Intelligent Information Systems* 46, 1 (2016), 83–97. https://doi.org/10.1007/s10844-014-0352-1 cited By 6.

[35] X. Li, X. Zhao, Z. Zhang, F. Wu, Y. Zhuang, J. Wang, and X. Li. 2016. Joint multilabel classification with community-aware label graph learning. *IEEE Transactions on Image Processing* 25, 1 (2016), 484–493. https://doi.org/10.1109/TIP.2015.2503700 cited By 14.

[36] Yaning Li and Liu Yang. 2021. More Correlations Better Performance: Fully Associative Networks for Multi-label Image Classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 9437–9444. https://doi.org/10.1109/ICPR48806.2021.9412004

[37] Zhitao Li, Jianzong Wang, Ning Cheng, and Jing Xiao. 2021. Semantic Embedding Graph Convolutional Networks for Multi-label Video Segment Classification. In *2021 12th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*. 146–151. https://doi.org/10.1109/PAAP54281.2021.9720457

[38] S.-M. Lian, J.-W. Liu, R.-K. Lu, and X.-L. Luo. 2019. Captured multi-label relations via joint deep supervised autoencoder. *Applied Soft Computing Journal* 74 (2019), 709–728. https://doi.org/10.1016/j.asoc.2018.10.035 cited By 2.

[39] Huiting Liu, Geng Chen, Peipei Li, Peng Zhao, and Xindong Wu. 2021. Multi-label text classification via joint learning from label embedding and label correlation. *NEUROCOMPUTING* 460 (OCT 14 2021), 385–398. https://doi.org/{10.1016/j.neucom.2021.07.031}

[40] H. Liu, Z. Wang, and Y. Sun. 2020. Stacking model of multi-label classification based on pruning strategies. *Neural Computing and Applications* 32, 22 (2020), 16763–16774. https://doi.org/10.1007/s00521-018-3888-0 cited By 2.

[41] Lin Liu and Lin Tang. 2018. A Survey of Statistical Topic Model for Multi-label Classification. In *2018 26TH INTERNATIONAL CONFERENCE ON GEOINFORMATICS (GEOINFORMATICS 2018) (International Conference on Geoinformatics)*, Hu, S and Ye, X and Yang, K and Fan, H (Ed.). Int Assoc Chinese Profess Geog Informat Sci; Wuhan Univ; Jiusan Soc, Yunnan Provincial Comm; SuperMap Software Co Ltd; Beijing PIESAT Informat Technol Co Ltd. 26th International Conference on Geoinformatics (Geoinformatics), Yunnan Normal Univ, Kunming, PEOPLES R CHINA, JUN 28-30, 2018.

[42] Shengyuan Liu, Haobo Wang, Tianlei Hu, and Ke Chen. 2021. Dual Enhancement for Multi-Label Learning with Missing Labels. In *2021 The 4th International Conference on Machine Learning and Machine Intelligence* (Hangzhou, China) *(MLMI'21)*. Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3490725.3490752

[43] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. 2021. The Emerging Trends of Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3119334

[44] Qiongdan Lou, Zhaohong Deng, Zhiyong Xiao, Kup-Sze Choi, and Shitong Wang. 2021. Multi-Label Takagi-Sugeno-Kang Fuzzy System. *IEEE Transactions on Fuzzy Systems* (2021), 1–1. https://doi.org/10.1109/TFUZZ.2021.3115967

[45] J. Ma, J. Fan, and W. Wang. 2017. Multi-label classification for images with missing labels. In *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*. 1050–1055. https://doi.org/10.1109/INDIN.2017.8104918

[46] J. Ma, Z. Tian, H. Zhang, and T.W.S. Chow. 2017. Multi-Label Low-dimensional Embedding with Missing Labels. *Knowledge-Based Systems* 137 (2017), 65–82. https://doi.org/10.1016/j.knosys.2017.09.005 cited By 8.

[47] M. Mei, Y. Zhong, F. He, and C. Xu. 2020. An innovative multi-label learning based algorithm for city data computing. *GeoInformatica* 24, 1 (2020), 221–245. https://doi.org/10.1007/s10707-019-00383-w cited By 0.

[48] J.M. Moyano, E.L. Gibaja, K.J. Cios, and S. Ventura. 2018. Review of ensembles of multi-label classifiers: Models, experimental study and prospects. *Information Fusion* 44 (2018), 33–45. https://doi.org/10.1016/j.inffus.2017.12.001 cited By 54.

[49] G. Nan, Q. Li, R. Dou, and J. Liu. 2018. Local positive and negative correlation-based k-labelsets for multi-label classification. *Neurocomputing* 318 (2018), 90–101. https://doi.org/10.1016/j.neucom.2018.08.035 cited By 6.

[50] Shabnam Nazmi, Abdollah Homaifar, and Mohd Anwar. 2021. An Effective Action Covering for Multi-Label Learning Classifier Systems: A Graph-Theoretic Approach. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Lille, France) *(GECCO '21)*. Association for Computing Machinery, New York, NY, USA, 340–348. https://doi.org/10.1145/3449639.3459372

[51] C. Nieuwenhuis, E. Töppe, and D. Cremers. 2013. A survey and comparison of discrete and continuous multi-label optimization approaches for the Potts model. *International Journal of Computer Vision* 104, 3 (2013), 223–240. https://doi.org/10.1007/s11263-013-0619-y cited By 39.

[52] R. C. Prati, F. Charte, and F. Herrera. 2017. A first approach towards a fuzzy decision tree for multilabel classification. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015521

[53] Niloofar Rastin, Mohammad Taheri, and Mansoor Zolghadri Jahromi. 2021. A stacking weighted k-Nearest neighbour with thresholding. *INFORMATION SCIENCES* 571 (SEP 2021), 605–622. https://doi.org/{10.1016/j.ins.2021.05.030}

[54] Reshma Rastogi and Sayed Mortaza. 2021. Multi-label classification with Missing Labels using Label Correlation and Robust Structural Learning. *KNOWLEDGE-BASED SYSTEMS* 229 (OCT 11 2021). https://doi.org/{10.1016/j.knosys.2021.107336}

[55] X. Shen, G. Dong, Y. Zheng, L. Lan, I. Tsang, and Q. Sun. 2021. Deep Co-Image-Label Hashing for Multi-label Image Retrieval. *IEEE Transactions on Multimedia* (2021). https://doi.org/10.1109/TMM.2021.3119868 cited By 0.

[56] M. Shi, Y. Tang, and X. Zhu. 2020. MLNE: Multi-Label Network Embedding. *IEEE Transactions on Neural Networks and Learning Systems* 31, 9 (Sep. 2020), 3682–3695. https://doi.org/10.1109/TNNLS.2019.2945869

[57] M. Shi, Y. Tang, X. Zhu, and J. Liu. 2020. Multi-Label Graph Convolutional Network Representation Learning. *IEEE Transactions on Big Data* (2020), 1–1. https://doi.org/10.1109/TBDATA.2020.3019478

[58] Wissam Siblini, Pascale Kuntz, and Frank Meyer. 2021. A Review on Dimensionality Reduction for Multi-Label Classification. *IEEE Transactions on Knowledge and Data Engineering* 33, 3 (March 2021), 839–857. https://doi.org/10.1109/TKDE.2019.2940014

[59] L. Sihao, C. Fucai, H. Ruiyang, and X. Yixi. 2017. Multi-label extreme learning machine based on label matrix factorization. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. 665–670. https://doi.org/10.1109/ICBDA.2017.8078719

[60] Y. Su, J. Xu, D. Hong, F. Fan, J. Zhang, and P. Jing. 2021. Deep low-rank matrix factorization with latent correlation estimation for micro-video multi-label classification. *Information Sciences* 575 (2021), 587–598. https://doi.org/10.1016/j.ins.2021.07.021 cited By 0.

[61] K. W. Sun, C. H. Lee, and J. Wang. 2016. Multilabel Classification via Co-Evolutionary Multilabel Hypernetwork. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (Sep. 2016), 2438–2451. https://doi.org/10.1109/TKDE.2016.2566621

[62] L. Sun, S. Feng, J. Liu, G. Lyu, and C. Lang. 2021. Global-Local Label Correlation for Partial Multi-Label Learning. *IEEE Transactions on Multimedia* (2021), 1–1. https://doi.org/10.1109/TMM.2021.3055959

[63] A.N. Tarekegn, M. Giacobini, and K. Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition* 118 (2021). https://doi.org/10.1016/j.patcog.2021.107965 cited By 9.

[64] J. Tu, G. Yu, C. Domeniconi, J. Wang, G. Xiao, and M. Guo. 2018. Multi-label Answer Aggregation Based on Joint Matrix Factorization. In *2018 IEEE International Conference on Data Mining (ICDM)*. 517–526. https://doi.org/10.1109/ICDM.2018.00067

[65] Jinzheng Tu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, and Maozu Guo. 2020. Multi-label crowd consensus via joint matrix factorization. *KNOWLEDGE AND INFORMATION SYSTEMS* 62, 4 (APR 2020), 1341–1369. https://doi.org/{10.1007/s10115-019-01386-7}

[66] Dengbao Wang, Fei Hu, and Li Li. 2017. Exploiting Label Correlations Using DBN Chains for Multi-Label Classification. In *Proceedings of the 12th Chinese Conference on Computer Supported Cooperative Work and Social Computing* (Chongqing, China) *(ChineseCSCW '17)*. Association for Computing Machinery, New York, NY, USA, 145–152. https://doi.org/10.1145/3127404.3127423

[67] H. Wang, Y. Zou, D. Chong, and W. Wang. 2020. Modeling Label Dependencies for Audio Tagging With Graph Convolutional Network. *IEEE Signal Processing Letters* 27 (2020), 1560–1564. https://doi.org/10.1109/LSP.2020.3019702

[68] K. Wang, M. Yang, W. Yang, and Y. Yin. 2018. Deep Cross-View Label Embedding with Correlation and Structure Preserved for Multi-Label Classification. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. 12–19. https://doi.org/10.1109/ICTAI.2018.00013

[69] R. Wang, R. Ridley, X. Su, W. Qu, and X. Dai. 2021. A novel reasoning mechanism for multi-label text classification. *Information Processing and Management* 58, 2 (2021). https://doi.org/10.1016/j.ipm.2020.102441 cited By 0.

[70] R. Wang, S. Ye, K. Li, and S. Kwong. 2021. Bayesian network based label correlation analysis for multi-label classifier chain. *Information Sciences* 554 (2021), 256–275. https://doi.org/10.1016/j.ins.2020.12.010 cited By 0.

[71] Ran Wang, Suhe Ye, Ke Li, and Sam Kwong. 2021. Bayesian network based label correlation analysis for multi-label classifier chain. *INFORMATION SCIENCES* 554 (APR 2021), 256–275. https://doi.org/{10.1016/j.ins.2020.12.010}

[72] S. Wang, G. Peng, and Z. Zheng. 2020. Capturing Joint Label Distribution for Multi-Label Classification through Adversarial Learning. *IEEE Transactions on Knowledge and Data Engineering* 32, 12 (2020), 2310–2321. https://doi.org/10.1109/TKDE.2019.2922603 cited By 0.

[73] Xueman Wang, Ling Du, and Junbing Li. 2021. Pmae: Pseudo Multi-Label Attention Ensemble. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME51207.2021.9428242

[74] Y. Wang, W. Zheng, Y. Cheng, and D. Zhao. 2021. Two-level label recovery-based label embedding for multi-label classification with missing labels. *Applied Soft Computing* 99 (2021). https://doi.org/10.1016/j.asoc.2020.106868 cited By 0.

[75] Z.-W. Wang, S.-K. Wang, B.-T. Wan, and W.W. Song. 2020. A novel multi-label classification algorithm based on K-nearest neighbor and random walk. *International Journal of Distributed Sensor Networks* 16, 3 (2020). https://doi.org/10.1177/1550147720911892 cited By 2.

[76] Tong Wei, Jiang-Xin Shi, and Yu-Feng Li. 2021. Probabilistic Label Tree for Streaming Multi-Label Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 1801–1811. https://doi.org/10.1145/3447548.3467226

[77] W. Weng, C.-L. Chen, S.-X. Wu, Y.-W. Li, and J. Wen. 2019. An Efficient Stacking Model of Multi-Label Classification Based on Pareto Optimum. *IEEE Access* 7 (2019), 127427–127437. https://doi.org/10.1109/ACCESS.2019.2931451_rfseq1 cited By 5.

[78] W. Weng, Y.-W. Li, J.-H. Liu, S.-X. Wu, and C.-L. Chen. 2021. Multi-label classification review and opportunities. *Journal of Network Intelligence* 6, 2 (2021), 255–275. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107979342&partnerID=40&md5=b482efe36008691685563f6ecbbe819b cited By 0.

[79] W. Weng, D.-H. Wang, C.-L. Chen, J. Wen, and S.-X. Wu. 2020. Label Specific Features-Based Classifier Chains for Multi-Label Classification. *IEEE Access* 8 (2020), 51265–51275. https://doi.org/10.1109/ACCESS.2020.2980551 cited By 0.

[80] G. Wu, R. Zheng, Y. Tian, and D. Liu. 2020. Joint Ranking SVM and Binary Relevance with robust Low-rank learning for multi-label classification. *Neural Networks* 122 (2020), 24–39. https://doi.org/10.1016/j.neunet.2019.10.002 cited By 9.

[81] J. Wu, A. Guo, V.S. Sheng, P. Zhao, and Z. Cui. 2018. An Active Learning Approach for Multi-Label Image Classification with Sample Noise. *International Journal of Pattern Recognition and Artificial Intelligence* 32, 3 (2018). https://doi.org/10.1142/S0218001418500052 cited By 5.

[82] Jian Wu, Anqian Guo, Victor S. Sheng, Pengpeng Zhao, Zhiming Cui, and Hua Li. 2017. Adaptive Low-Rank Multi-Label Active Learning for Image Classification. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 1336–1344. https://doi.org/10.1145/3123266.3123388

[83] J. Wu, S. Ruan, C. Lian, S. Mutic, M. A. Anastasio, and H. Li. 2018. Active learning with noise modeling for medical image annotation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 298–301. https://doi.org/10.1109/ISBI.2018.8363578

[84] J. Wu, C. Ye, V.S. Sheng, J. Zhang, P. Zhao, and Z. Cui. 2017. Active learning with label correlation exploration for multi-label image classification. *IET Computer Vision* 11, 7 (2017), 577–584. https://doi.org/10.1049/iet-cvi.2016.0243 cited By 13.

[85] J. Wu, S. Zhao, V.S. Sheng, J. Zhang, C. Ye, P. Zhao, and Z. Cui. 2017. Weak-Labeled Active Learning With Conditional Label Dependence for Multilabel Image Classification. *IEEE Transactions on Multimedia* 19, 6 (2017), 1156–1169. https://doi.org/10.1109/TMM.2017.2652065 cited By 20.

[86] J. Wu, S. Zhao, V. S. Sheng, P. Zhao, and Z. Cui. 2016. Multi-label active learning for image classification with asymmetrical conditional dependence. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2016.7552899

[87] Q. Wu, M. Tan, H. Song, J. Chen, and M.K. Ng. 2016. ML-FOREST: A multi-label tree ensemble method for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2665–2680. https://doi.org/10.1109/TKDE.2016.2581161 cited By 48.

[88] Yuelong Xia, Ke Chen, and Yun Yang. 2021. Multi-label classification with weighted classifier selection and stacked ensemble. *INFORMATION SCIENCES* 557 (MAY 2021), 421–442. https://doi.org/{10.1016/j.ins.2020.06.017}

[89] Y. Xie, X. Wang, D. Jiang, X. Xu, G. Bao, and R. Xu. 2018. Multi-label Green's Function Criterion inspired Transfer Annotation System. In *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*. 737–742. https://doi.org/10.1109/CSCWD.2018.8465314

[90] C. Xu, T. Liu, D. Tao, and C. Xu. 2016. Local Rademacher Complexity for Multi-Label Learning. *IEEE Transactions on Image Processing* 25, 3 (March 2016), 1495–1507. https://doi.org/10.1109/TIP.2016.2524207

[91] Z. Xu, Y. Liu, and C. Li. 2020. Distributed Information-Theoretic Semisupervised Learning for Multilabel Classification. *IEEE Transactions on Cybernetics* (2020), 1–15. https://doi.org/10.1109/TCYB.2020.2986463

[92] L. Xue, D. Jiang, R. Wang, J. Yang, and M. Hu. 2020. Learning semantic dependencies with channel correlation for multi-label classification. *Visual Computer* 36, 7 (2020), 1325–1335. https://doi.org/10.1007/s00371-019-01731-5 cited By 0.

[93] Y. Yan, S. Li, and L. Feng. 2021. Partial multi-label learning with mutual teaching. *Knowledge-Based Systems* 212 (2021). https://doi.org/10.1016/j.knosys.2020.106624 cited By 0.

[94] J. Yao, K. Wang, and J. Yan. 2019. Incorporating Label Co-Occurrence Into Neural Network-Based Models for Multi-Label Text Classification. *IEEE Access* 7 (2019), 183580–183588. https://doi.org/10.1109/ACCESS.2019.2960626

[95] L. Yao, Q.Z. Sheng, A.H.H. Ngu, B.J. Gao, X. Li, and S. Wang. 2016. Multi-label classification via learning a unified object-label graph with sparse representation. *World Wide Web* 19, 6 (2016), 1125–1149. https://doi.org/10.1007/s11280-015-0376-7 cited By 4.

[96] Y. Yao, Y. Li, Y. Ye, and X. Li. 2021. MLCE: A Multi-Label Crotch Ensemble Method for Multi-Label Classification. *International Journal of Pattern Recognition and Artificial Intelligence* 35, 4 (2021). https://doi.org/10.1142/S021800142151006X cited By 1.

[97] P. Ye, S. Feng, H. Feng, and G. Dai. 2019. Robust Multi-Label Learning with Corrupted Features and Incomplete Labels. In *2019 Chinese Automation Congress (CAC)*. 4411–4416. https://doi.org/10.1109/CAC48633.2019.8996261

[98] T. Yu, G. Yu, J. Wang, C. Domeniconi, and X. Zhang. 2020. Partial Multi-label Learning using Label Compression. In *2020 IEEE International Conference on Data Mining (ICDM)*. 761–770. https://doi.org/10.1109/ICDM50108.2020.00085

[99] J. Zhang, Z. He, J. Zhang, and T. Dai. 2019. Cograph Regularized Collective Nonnegative Matrix Factorization for Multilabel Image Annotation. *IEEE Access* 7 (2019), 88338–88356. https://doi.org/10.1109/ACCESS.2019.2925891

[100] J. Zhang, S. Li, M. Jiang, and K. C. Tan. 2020. Learning From Weakly Labeled Data Based on Manifold Regularized Sparse Model. *IEEE Transactions on Cybernetics* (2020), 1–14. https://doi.org/10.1109/TCYB.2020.3015269

[101] J.-J. Zhang, M. Fang, and X. Li. 2017. Clustered intrinsic label correlations for multi-label classification. *Expert Systems with Applications* 81 (2017), 134–146. https://doi.org/10.1016/j.eswa.2017.03.054 cited By 3.

[102] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* 26, 8 (AUG 2014), 1819–1837. https://doi.org/{10.1109/TKDE.2013.39}

[103] Q. Zhang, G. Zhong, and J. Dong. 2021. A Graph-based Semi-supervised Multi-label Learning Method Based on Label Correlation Consistency. *Cognitive Computation* 13, 6 (2021), 1564–1573. https://doi.org/10.1007/s12559-021-09912-y cited By 0.

[104] X. Zhang, W. Li, H. Ying, F. Li, S. Tang, and S. Lu. 2020. Emotion Detection in Online Social Networks: A Multilabel Learning Approach. *IEEE Internet of Things Journal* 7, 9 (2020), 8133–8143. https://doi.org/10.1109/JIOT.2020.3004376 cited By 0.

[105] Xiulin Zheng, Peipei Li, Zhe Chu, and Xuegang Hu. 2020. A Survey on Multi-Label Data Stream Classification. *IEEE Access* 8 (2020), 1249–1275. https://doi.org/10.1109/ACCESS.2019.2962059

[106] T. Zhong, F. Liu, F. Zhou, G. Trajcevski, and K. Zhang. 2019. Motion Based Inference of Social Circles via Self-Attention and Contextualized Embedding. *IEEE Access* 7 (2019), 61934–61948. https://doi.org/10.1109/ACCESS.2019.2915535

A SEARCH STRINGS USED IN THE SCIENTIFIC REPOSITORIES

ACM Digital Library

[[[Publication Title: multilabel] OR [Publication Title: "multi label"] OR [Publication Title: "multilabel"]] AND [[Publication Title: classification] OR [Publication Title: learning]] AND [[Publication Title: "label correlation"] OR [Publication Title: "label correlations"] OR [Publication Title: "label dependenc*"] OR [Publication Title: "correlation* among labels"] OR [Publication Title: "correlation* between labels"] OR [Publication Title: "dependenc* among labels"] OR [Publication Title: "relation* among the labels"] OR [Publication Title: "label relation*"]]] OR [[[Abstract: multilabel] OR [Abstract: "multi label"] OR [Abstract: "multilabel"]] AND [[Abstract: classification] OR [Abstract: learning]] AND [[Abstract: "label correlation"] OR [Abstract: "label correlations"] OR [Abstract: "label dependenc*"] OR [Abstract: "correlation* among labels"] OR [Abstract: "correlation* between labels"] OR [Abstract: "dependenc* among labels"] OR [Abstract: "relation* among the labels"] OR [Abstract: "label relation*"]]] OR [[[Keywords: multilabel] OR [Keywords: "multi label"] OR [Keywords: "multilabel"]] AND [[Keywords: classification] OR [Keywords: learning]] AND [[Keywords: "label correlation"] OR [Keywords: "label correlations"] OR [Keywords: "label dependenc*"] OR [Keywords: "correlation* among labels"] OR [Keywords: "correlation* between labels"] OR [Keywords: "dependenc* among labels"] OR [Keywords: "relation* among the labels"] OR [Keywords: "label relation*"]]] AND [Publication Date: (01/01/2016 TO *)]

Scopus

TITLE-ABS-KEY (multilabel OR "multi label" OR "multilabel") AND TITLE-ABS-KEY (classification OR learning) AND TITLE-ABS-KEY ("label correlation" OR "label correlations" OR "label dependenc*" OR "correlation* among labels" OR "correlation* between labels" OR "dependenc*

among labels" OR "relation* among the labels" OR "label relation*") AND PUBYEAR > 2015 AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI")) AND (LIMIT-TO (DOCTYPE , "ar"))

IEEE Xplore

(("All Metadata":multilabel OR "multi label" OR "multilabel") AND ("All Metadata":classification OR learning) AND ("All Metadata":"label correlation" OR "label correlations" OR "label dependenc*" OR "correlation* among labels" OR "correlation* between labels" OR "dependenc* among labels" OR "relation* among the labels" OR "label relation*"))

Web of Science

"TOPIC: (multilabel OR ""multi label"" OR ""multilabel"") AND TOPIC: (classification OR learning) AND TOPIC: (""label correlation"" OR ""label correlations"" OR ""label dependenc*"" OR ""correlation* among labels"" OR ""correlation* between labels"" OR ""dependenc* among labels"" OR ""relation* among the labels"" OR ""label relation*"") Refined by: WEB OF SCIENCE CATEGORIES: (COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR ENGINEERING ELECTRICAL ELECTRONIC OR COMPUTER SCIENCE INFORMATION SYSTEMS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS) AND DOCUMENT TYPES: (ARTICLE) Timespan: 2016-2021."