International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-022-01658-9

ORIGINAL ARTICLE

# A survey of multi-label classification based on supervised and semi-supervised learning

**Meng Han[1] · Hongxin Wu[1] · Zhiqiang Chen[1] · Muhang Li[1] · Xilong Zhang[1]**

Received: 24 September 2021 / Accepted: 11 September 2022


**Abstract**
Multi-label classification algorithms based on supervised learning use all the labeled data to train classifiers. However, in real life, many of the data are unlabeled, and it is costly to label all the data needed. Multi-label classification algorithms based on semi-supervised learning can use both labeled and unlabeled data to train classifiers, resulting in better-performing models. In this paper, we first review supervised learning classification algorithms in terms of label non-correlation and label correlation and semi-supervised learning classification algorithms in terms of inductive methods and transductive methods. After that, multi-label classification algorithms are introduced from the application areas of image, text, music and video. Subsequently, evaluation metrics and datasets are briefly introduced. Finally, research directions in complex concept drift, label complex correlation, feature selection and class imbalance are presented.

**Keywords** Supervised learning · Semi-supervised learning · Image classification · Text classification · Evaluation metrics

## 1 Introduction

With the rapid development of big data, a large amount of data is generated in life, and these data contain a lot of information closely related to human life. In order to obtain the required data, many tasks related to data mining have been carried out [1]. Traditional classification methods focus on single-label classification. However, many practical problems require multi-label classification (MLC). The goal of MLC is to predict the potential multiple labels of the test set by a prediction model based on the training set [2].

The classic MLC methods are mainly divided into problem transformation (PT) and algorithm adaptation problem (AA). The most commonly used in PT is the Binary Relevance (BR) method. The BR method does not consider the interdependence between labels. In order to overcome this problem, researchers proposed the classifier chains method (CC) [3], which is based on BR and connects the binary classifier obtained by BR through a chain method. The label power-set (LP) method is also PT. The RAndom k-labELsets (RAkEL) [4] is an ensemble use of LP, where each LP classifier is trained by a different small subset of randomly generated labels. AA is the modification of an existing algorithm to fit the new problem to be solved. The specific performance is to adjust the existing single-label classification problem to the MLC problem. Popular models of AA for building multi-label classifiers include *k*-Nearest Neighbor (kNN) [5], decision tree [6], Support Vector Machine (SVM) [7], Neural Networks (NN) [8] and so on.

In recent years, several surveys on MLC have been provided. Tsoumakas et al. [9] detailed MLC methods from the perspective of PT and AA, briefly introduced some evaluation metrics, and finally compared the experimental results of MLC methods. Moyano et al. [10] compared multi-label ensemble classifiers on 20 datasets and evaluated their performance based on the characteristics of imbalanced datasets and the correlation between labels. Zheng et al. [11] introduced traditional MLC methods and multi-label data stream classification algorithms from multi-label data stream classification, discussed their advantages and disadvantages, and determined the mining constraints of multi-label data stream classification. Sadarangani et al. [12] only introduced semi-supervised learning from the perspective of single label. Supervised learning (SL) is one of the branches of machine learning, which can be divided into regression and classification problems. Semi-supervised learning (SSL) is a popular method to deal with incomplete markings. So far, no survey

✉ Meng Han
2003051@nmu.edu.cn

[1] School of Computer Science and Engineering, North Minzu University, Yinchuan, China

Published online: 28 October 2022

has introduced MLC from the perspective of SL and SSL, and no survey has provided a comprehensive introduction to the practical application of multi-label. The overall framework of this article is shown in Fig. 1.

The main contributions of this paper are:

(1) We present a comprehensive review of MLC algorithms based on SL and SSL and summarize the existing MLC algorithms and discuss their advantages and disadvantages.
(2) We have studied and summarized the MLC algorithms from the perspective of application fields.
(3) We introduce the commonly used evaluation metrics and open datasets, and graphically demonstrate the evaluation metrics used by SL classification algorithms.
(4) We analyze the problems in MLC algorithms, and propose the next research directions.

The remainder of this paper is organized as follows: Section 2 describes MLC based on SL and SSL. Section 3 describes application fields, including image classification, text classification, and other fields. Section 4 mainly introduces the evaluation metrics and datasets. Finally, Sects. 5 and 6 propose further research directions and conclude the whole paper, respectively.

## 2 Multi-label classification based on supervised and semi-supervised learning

Both supervised and semi-supervised learning algorithms have been widely used in multi-label classification. This section will summarize them from the perspective of supervised and semi-supervised learning.

### 2.1 Supervised learning

SL is a machine learning task that infers a function from a labeled training set [13]. In the next few sections, the paper will review MLC algorithms from two aspects: SL based on non-label correlation and SL based on label correlation.

#### 2.1.1 Label non-correlation algorithm

In MLC, the correlation between the labels is very complex [14]. Without considering this problem, the difficulty of the algorithm can be simplified. Based on label non-correlation, this section introduces MLC algorithms from multiple directions such as decision trees, Bayes, SVM, NN, kNN and ensemble.

Multi-label decision trees for prediction probability [15] build a tree using a traditional, single-label decision tree algorithm in the context of SL, using a normalization method to convert multi-label data into single-labeled instances. The algorithm evaluates the method based on the performance of tree complexity and prediction accuracy, introducing a new metric for comparison of datasets. LdSM [16] can be used to build and train multi-label decision trees with a new objective function optimized in each node of the tree that facilitates balanced splitting, maintains high-class purity of the child nodes, and allows sending instances in multiple directions with penalties to prevent excessive tree overgrowth. Once the previous node is completed, each node of the tree is trained. ML-decision trees based on NPI-M [17] is a new nonparametric predictive inference model based on multinomial data, and the splitting criterion of this algorithm makes it independent of the noise of the labels, and the imprecise information gain is calculated as follows, where $H^*(L|A = a_i)$ is the maximum value of $H^*(L)$ in the partition of the dataset consisting of instances with $A = a_i$.

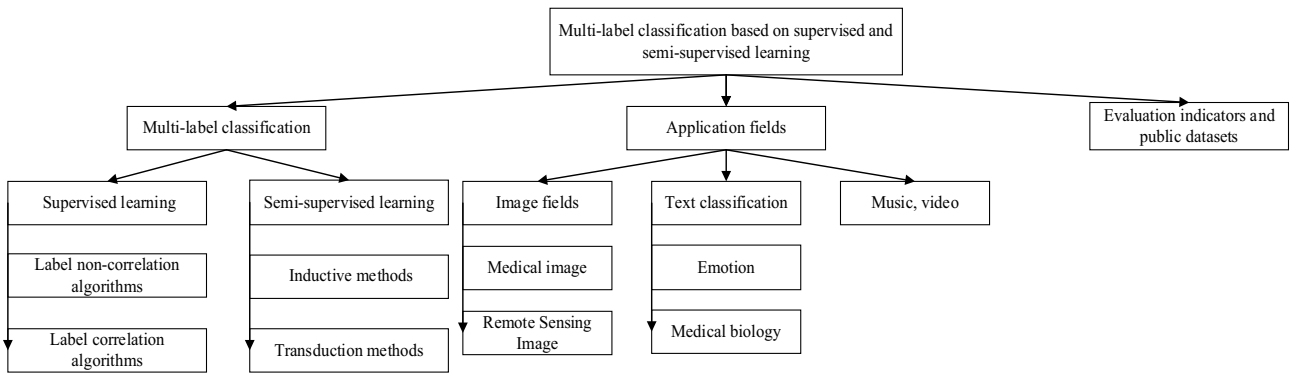


**Fig. 1** Overall framework diagram

$$IIG(L,A) = H^*(L) - \sum_{i=1}^{n} P\big(A = a_i\big) H^*(L|A = a_i) \quad (1)$$

Using Bayesian networks as base models, Yang et al. [18] proposed a feature weighting method to improve the classification accuracy of the decision function. The conditional probabilities of positive classes are estimated by computing the frequency ratios of features in-depth from the training data, and the decision function can be simplified by eliminating redundant variables for variables whose probabilities are independent of the decision function.

AEML-LLSVM [19] is a fast classification algorithm based on multi-core low-rank-linearized SVM. The BR transformation strategy is used to decompose the multi-label data set into multiple binary data sets. Then, the approximate extreme value method is used to obtain a representative set from each binary data set. Finally, the algorithm is trained on each representative set to achieve MLC. AEDC-MLSVM [20] is an algorithm that combines approximate extreme value method and divide-and-conquer strategy under SL, and is an improved approach to the previous algorithm. The algorithm also uses BR to implement MLC and is suitable for use with large-scale datasets with relatively low time complexity.

Under SL, BP-AEPML [21] uses a method to reduce the size of the original dataset by extracting a representative dataset based on approximate limit points, and then a BP neural network is used to train the representative dataset. NNMLInf [22] is a prediction model based on NN. This model can be used to predict social influence. Among them, people's network structural features are considered network inputs and their behaviors are classified into multiple labels as network outputs. The algorithm of deep neural architecture based on bidirectional correlation pool layer [23] uses a correlation function to detect different pairs of neurons that will be aggregated into merged neurons. An iterative procedure is proposed, which can estimate the correlation between the merged neurons in the deeper layer without recomputing the correlation matrix. A novel multi-attention drive system for remote sensing [24] proposes a $k$-branch CNN to extract the preliminary local descriptors of remote sensing image bands associated with different spatial resolutions. All the outputs of the RNN are used to predict the multi-label of remote sensing images, instead of determining each label by considering a single class-specific node.

Extreme Learning Machine (ELM) is a method based on feedforward NN construction and is used in MLC due to its fast training [25]. ML-KELM [26] solves the problem of converting the real-valued output of the network to a binary vector using an adaptive threshold function, while adjusting fewer parameters, running stably, converging fast and generalizing well. ML-CK-ELM [27] uses linearly combined basis kernels in each layer, which does not require random adjustment of parameters and has a significant reduction in computation time and memory storage. Rr et al. [28] proposed two frameworks, RMLFM applies the feature manifold regularization term and RMLDM considers both feature manifold and data manifold regularization to maintain the local structure of data and features, while two iterative algorithms based on the global conjugate gradient method are used to solve the objective functions of the proposed methods RMLFM and RMLDM.

ML-RkNN [29] is a neighbor-based reverse nearest neighbor MLC algorithm. For the same value of $k$, the algorithm adaptively acquires different numbers of neighbors for different instances, thus better learning the local configurations around the points. Also, by comparing the class distribution of test points and their reverse nearest neighbors, it helps to implicitly deal with the local imbalance problem prevalent in the dataset. To address the data stream problem, MLSAMPkNN [30] uses self-tuning memory to accommodate various types of concept drift, implementing a penalty system to identify and remove instances of introduced errors. Instances that have a significant impact on the error are quickly removed, the fact that it was recently added to the window. By removing poorly performing instances, punitive systems can also help keep memory sizes small and reduce the amount of computation required to determine the distance of incoming instances. MLSAkNN [31] uses the penalties of MLSAMPkNN, but it assumes that the instance causing the error is completely wrong, but it is possible that an instance has a noisy label or that conceptual drift affects only a few labels. The algorithm proposes methods to enable and disable dynamic instances and instances of each label.

The ELIFT [32] based on SL uses an ensemble method to alleviate the limitation on high classification accuracy. Multiple training sets generated using a bagging strategy are used to construct multiple LIFT classifiers. According to the loss of each classifier, different classifiers are automatically weighted. For each new instance, the predicted label vector is obtained through the learned weighted ensemble classifier. AESAKNNS [33] uses MLSAkNN as the base classifier to take advantage of ensemble classification to accommodate concept drift in multi-label environment. The ADWIN detector monitors each classifier for concept drift on a subspace. Once detected, the algorithm automatically trains additional classifiers in the background to try to capture new concepts on new feature subspaces. The dynamic classifier selects the most accurate classifier from the active and background ensemble to replace the current ensemble.

#### 2.1.2 Label correlation algorithm

In many practical tasks, labels are highly correlated, so the key to successful multi-label learning is to effectively utilize the correlation between different labels [34]. This section introduces many aspects such as decision trees, Bayes, SVM, NN and ensemble algorithms.

In the case of SL, learning the correlation of labels may produce circular dependence. To solve this problem, the 3RC [35] is proposed. This new method follows the BR method and uses multiple decision trees as binary classifiers. This novel method aims to learn the correlation of labels and gives results for models that only consider relevant dependencies in order to perform better predictions and reduce error propagation due to irrelevant and weak dependencies. ML-BTC [36] is an extended algorithm for decision trees in which a new labeled space partitioning technique is applied to the data to implicitly handle the possibility of being overlooked in the process of constructing the tree potential class associations. The tree is constructed based on parameters that act as restrictions to prevent unnecessary branching for smaller imbalanced classes.

MLNB-LD [37] proposes Bayes' theorem with strong independence hypothesis, a new posterior probability estimation method for multi-label problems. The proposed method uses the correlations between label pairs to determine the most likely label set for a given unseen instance. BCC considers the importance of feature selection in classification tasks. To improve the performance of classification by improving each internal classifier, two algorithms are proposed to test BCC [38], namely BF-FS-BCC and GS-FS-BCC. Given the structure and chain sequence of BCC, for each label, a subset is selected and a classifier is built. BNCC [39] uses conditional entropy to describe the relationship between labels, with nodes as labels and the weights of the edges as associations to construct the BN. It proposes a scoring function to evaluate the BN structure and introduces a heuristic algorithm to optimize the BN structure, and derives the label order for constructing the CC model by topologically ranking the nodes of the optimized Bayesian network.

Under SL, RBRL [40] is an algorithm that combines ranking SVM, BR and robust low-rank learning. It captures the nonlinear relationship between input and output, and uses two accelerated approximate gradient algorithms. The accelerated proximal gradient method (APG) to effectively solves the fast converging optimization problem. SSSVM [41] is an SVM method for ultra-high resolution remote sensing images. The basic idea is to exploit the relationship between labels through structured SVM and to incorporate spatial background information into the structured SVM optimization process by adding terms to the cost function that encourage spatial smoothing.

LCL-Net [42] introduces a multilayer perceptron into the LCL module to model the correlation between conditions in the ChestX-ray14 dataset. The multilayer perceptron is a general function approximator, which can adaptively recalibrate the multi-label output during the training phase to improve the performance of LCL-Net. At the same time, the LCL module can be easily inserted at the end of any CNN-based model. SDLM [43] model uses a convolutional neural network called VGG to learn whole brain CT images in the image feature learning part. Under SL, the slice correlation between variable-length slices and the causal relationship between multiple diseases can be obtained from RNN. DCNet [44] is mainly composed of three main modules. Among them, feature extraction is the backbone CNN, which is used for the spatial correlation model of feature association and the classifier used for classification score generation. The features generated by the backbone CNN are directly fused through summation, pixel-by-pixel multiplication or cascading, without a special fusion process. Krishna and Prakash [45] used multiple convolutional layers to form a deep neural network and extract features at different levels. The classifier learns from previously unknown trends while discovering potential dependencies between labels. Zhou et al. [46] used to structure, attribute and label information to solve the multi-label graph node classification problem. The model uses the one-dimensional convolution operator of TextCNN to extract node feature representations while embedding the nodes into the same vector space. The dimensionality of the feature representation learned by the algorithm is independent of the size of the node neighborhood. It uses an additional attention mechanism to measure the compatibility of node labels.

Each output node of ML-RBF is connected to the output of the hidden layer, and the correlation between different classes can be properly handled to obtain the output weights. Nan et al. [47] proposed two methods, WuELM-AE and ML-ELM-RBF, respectively. WuELM-AE introduces the uncertainty of weights and treats the input weights as random variables obeying Gaussian distribution. ML-ELM- RBF first overlays WuELM-AE, then performs cluster analysis on the sample features of each possible class, and finally uses a regularized least squares resolution to calculate the output weights of ML-ELM-RBF. ML-AP-RBF-Lap-ELM [25] uses ML-RBF for mapping at the input layer, and the affinity propagation clustering algorithm can automatically determine the number and center of hidden nodes of the RBF function and use Lap-ELM to solve the weights from the hidden layer to the output layer.

ACkEL [48] is an ensemble classification method, which borrows the idea of active learning and proposes label selection criteria to evaluate the separability and balance level of classes transformed from a labeled subset. The $K$-label ensemble method based on mutual information and joint

**Table 1** Table of time complexity analysis

| Algorithm | Time complexity |
|---|---|
| AEML-LLSVM [19] | $O(LM^2/k)$ |
| AEDC-MLSVM [20] | Training time complexity: $O(LMm^2)$ |
| BP-AEPML [21] | $O(nuM)$ |
| ML-RkNN [29] | In general: $O(N^2+d)$ |
| MLSAMPkNN [30] | Average time complexity: $\mathrm{O}\left(wlog_2\frac{w}{w_{min}}\right)$, Worst case scenario: $\mathrm{O}\left(w_{max}log_2\frac{w_{max}}{w_{min}}\right)$ |
| MLSAkNN [31] | Average time complexity: $\mathrm{O}\left(wd+wL+wlog_2\frac{w}{w_{min}}\right)$, Worst case scenario: $\mathrm{O}\left(w_{max}d+w_{max}L+w_{max}log_2\frac{w_{max}}{w_{min}}\right)$ |

entropy [49] evaluates the redundancy and imbalance of each $K$-label set. The algorithm iteratively performs discrete sampling, retains multiple $K$-label sets with low mutual information as candidates, and selects the $K$-label set with the highest joint entropy.

### 2.1.3 Summary

This section provides a detailed description of some algorithms for label non-correlation and labels correlation from four perspectives: time complexity, experimental results, technical methods and performance.

**2.1.3.1 Analysis of time complexity** Generally speaking, compared to decision trees, Bayesian and ensemble algorithms, SVM have limited their application in large-scale data sets due to time complexity issues. The algorithm in this section effectively solves this problem. Among them, RBRL [40] and AEDC-MLSVM [20] can solve the problem of class imbalance. The neural network also has the problem of too long training time. The learning rate of the BP neural network MLC algorithm is fixed. To make the output vector as close to the expected value as possible, the weight and bias of the network need to be adjusted repeatedly during the data training process. In this process, the larger the size of the training data set, the longer the adjustment time required, especially when there are more hidden layers. Extracting representative data sets based on approximate limit points can reduce the size of the original data set and reduce the time spent on data training.

Table 1 shows the complexity analysis of the MLC algorithms for SVM and KNN under SL. $L$ denotes the number of labels, $M$ denotes the representative size, $k$ represents the number of cluster centers, n denotes the number of hidden layers, m denotes the number of landmark data instances, $u$ denotes the number of cells in each hidden layer, $d$ denotes the feature vector, $N$ denotes the training set cardinality, $w$ denotes the window size. As can be seen in Table 1, AEML-LLSVM has lower time complexity than AEDC-MLSVM, while MLSAkNN sacrifices time at the cost of improving classification results.

**Table 2** Table of ELM result analysis

| Algorithm | Yeast | Scene |
|---|---|---|
| ML-AP-RBF-Lap-ELM [25] | 0.7673 | 0.8121 |
| ML-KELM [26] | 0.7702 | **0.8850** |
| ML-CK-ELM [27] | 0.7702 | 0.8148 |
| ML-ELM-RBF [47] | **0.7673** | 0.8299 |

The bolded values indicate the highest results obtained in the corresponding data set in table

**Table 3** Table of kNN result analysis

| Algorithm | Mediamill | Imdb | Nuswide-C |
|---|---|---|---|
| MLSAMPkNN [30] | 0.160 | 0.066 | 0.247 |
| MLSAkNN [31] | 0.152 | 0.072 | **0.251** |
| AESAKNNS [33] | **0.189** | **0.093** | 0.240 |

The bolded values indicate the highest results obtained in the corresponding data set in table

**2.1.3.2 Analysis of experimental results** Both NNMLInf and BP-AEPML were compared with SVM and both improved in terms of time efficiency. On the dataset Mediamill, BP-AEPML has an average precision of 71.12% and takes 622.7 s when the number of hidden layers is 5, but SVM is only 38.99% and takes 13371 s. BP-AEPML uses approximate extreme points to extract the representative set and the size of the representative set is smaller than the original set. Therefore, the time to adjust the weights and thresholds is reduced. Then the training time is reduced. The time complexity of the SVM is $M^3$, and $M$ is the size of the dataset, making its running time slower. On the dataset Youtube, NNMLInf has an average precision of 46% and SVM of 41.85% when the number of hidden layers is 5.

Table 2 shows the average precision of ELM algorithms. The algorithms in the table are experimented with using five-fold cross-validation. ML-KELM is set with an adaptive threshold function, which gives it a faster convergence and better generalization performance. On the dataset yeast, the running time of ML-KELM is 0.56 s, which is 1.3276 s

faster than ML-ELM-RBF. Meanwhile, in the larger dataset Delicious, ML-ELM-RBF can obtain a better average accuracy of 38.2%, compared with 37.73% for ML-CK-ELM and ML-ELM-FM-DM is 35.74%, but ML-CK-ELM is in a better result except the AP and error are lower than ML-ELM-RBF, while the Coverage of ML-CK-ELM is 547.0185, which is smaller than both ML-ELM-RBF.

Table 3 shows the subset accuracy of kNN algorithms. For a summary of kNN algorithms on data streams, which can all cope with the concept drift problem, the table only records the experimental results on larger data sets. AESAKNNS can overcome various multi-label data difficulties due to its combinatorial mechanism.

**2.1.3.3 The technical analysis involved in algorithms** Table 4 summarizes the classification methods, and the correlation between labels, and deals with imbalances between classes mentioned in the paper. BF-FS-BCC is an extended algorithm of BCC [38], which considers the correlation between labels, and experiments with the BR algorithm, which does not process the label relationships, reveal that BF-FS-BCC can obtain good classification results in terms of Hamming score, accuracy and Macro accuracy. Among them, in the dataset Medical, the accuracy of this algorithm is 72.5%, which is 18.4% higher than BR. The class imbalance problem is more interesting and challenging for multi-label datasets [29].ML-BTC and ML-RkNN consider the problem of class imbalance and the classification effect is effectively improved. On the dataset CHD49, the Macro F1 value of the comparison algorithm MK-KNN is 40.92% and ML-BTC is 43.38%. On the dataset Yeast, the Macro F1 value of the comparison algorithm MLkNN is 37.82%, while that of ML-RkNN is 45.28%.

**2.1.3.4 Analysis of performance of algorithms** To facilitate the analysis of performance of algorithms, Table 5 summarizes the SL algorithms mentioned in the paper in terms of comparing algorithms, experimental datasets, testing domains, and advantages and disadvantages. In general, BR is widely used in many fields because of its simple implementation and fast running speed, but it ignores the relationship between labels and treats each label separately, losing

**Table 4** Table of technical analysis

| Algorithm | Classification method | Correlation between labels | Deal with imbalances between classes |
|---|---|---|---|
| LdSM [16] | Decision trees | No | No |
| 3RC [35] | Decision trees | Yes | No |
| ML-BTC [36] | Decision trees | Yes | Yes |
| BCC [38] | Bayes | Yes | No |
| BNCC [39] | Bayes | Yes | No |
| RBRL [40] | SVM | Yes | No |
| AEDC-MLSVM [20] | SVM | No | Yes |
| SSSVM [41] | SVM | Yes | No |
| AEML-LLSVM [19] | SVM | No | No |
| BP-AEPML [21] | NN | No | No |
| NNMLInf [22] | NN | No | No |
| LCL-Net [42] | NN | Yes | Yes |
| SDLM [43] | NN | Yes | No |
| DCNet [44] | NN | Yes | No |
| LANC [46] | NN | Yes | No |
| ML-ELM-RBF [47] | ELM | Yes | No |
| ML-ELM-FM-DM [28] | ELM | No | No |
| ML-CK-ELM [27] | ELM | No | No |
| ML-KELM [26] | ELM | No | No |
| ML-AP-RBF-Lap-ELM [25] | ELM | Yes | No |
| ML-RkNN [29] | KNN | No | Yes |
| MLSAMPkNN [30] | KNN | No | Yes |
| MLSAkNN [31] | KNN | No | Yes |
| AESAKNNS [33] | Ensemble | No | Yes |
| ELIFT [32] | Ensemble | No | No |
| ACkEL [48] | Ensemble | Yes | Yes |

**Table 5** Table of classification performance

| Algorithm | Comparison algorithms | Experimental datasets | Test fields | Advantages and disadvantages |
|---|---|---|---|---|
| LdSM [16] | GBDT-S; CRAFTML; FastXML; PFastreXML | Bibtex; Mediamill; Delicious; AmazonCat-13 k; Wiki10-31 k; etc | Text; Video | *Advantages:* The algorithm is suitable for applications with large label spaces<br>*Disadvantages:* The algorithm does not exploit the correlation between labels to improve the classification performance |
| 3RC [35] | BR; RAKEL; MLKNN | Scene; Yeast; emotions; Mediamill | Image; Biology; Music; Text | *Advantages:* The algorithm introduces dependency weights to determine their relevance and retains only the important dependencies<br>*Disadvantages:* The algorithm has not been experimented with in a big data environment |
| MLNB-LD [36] | MLNB; MLDT; ML-kELM; GLOCAL | Arts; Education; Entertain; Health; Recreation; Reference; Science; Social; etc | Text; Images; Music | *Advantages:* The algorithm uses the correlation between label pairs to determine the most likely set of labels<br>*Disadvantages:* The algorithm has low computational efficiency and does not remove unnecessary noisy information |
| RBRL [40] | Rank-SVM; Rank-SVMz; BR;ML-kNN; CLR; RAKEL; etc | Emotions; image; scene; yeast; enron; arts; education; recreation; etc | Music; Images; Biology; Text; | *Advantages:* RBRL can reduce hamming loss and sorting loss at the same time<br>*Disadvantages:* It does not consider the performance in the case of unbalanced labels |
| AEDC-MLSVM [20] | ML-LIBSVM; ML-CVM; ML-BVM | TMC2007-500; Mediamill; EukaryoteGO | Text; Audio; Biology; Image | *Advantages:* The training and testing time of the algorithm is greatly reduced<br>*Disadvantages:* The algorithm does not use the correlation between the labels to improve the classification performance |
| SSSVM [41] | SVM; SSVM | UAV data set; airborne image data set | Image | *Advantages:* This method can model spatial continuity and output structure<br>*Disadvantages:* Because each step requires solving a linear programming problem for all samples, it is computationally expensive |
| AEML-LLSVM [19] | ML-LIBSVM; ML-CVM; ML-BVM | Mediamil; Tmc2007-500; Nuc-wide | Audio; Text | *Advantages:* The training and testing time is greatly shortened under the premise of maintaining the performance advantage<br>*Disadvantages:* The correlation between labels is not considered |

**Table 5** (continued)

| Algorithm | Comparison algorithms | Experimental datasets | Test fields | Advantages and disadvantages |
|---|---|---|---|---|
| BP-AEPML [21] | Basic BP; ML-BVM; SVM | Mediamill | Video | *Advantages:* The training time of the algorithm is greatly reduced<br>*Disadvantages:* The algorithm is not considered for experimentation on multiple application domains |
| SDLM [43] | 3D-VGG; C3D | CQ500 data set | Medical; Image | *Advantages:* The model can learn the features of sequential images and their slice dependence in whole brain CT scans<br>*Disadvantages:* The algorithm does not perform image enhancement for continuous images |
| DCNet [44] | ResNet-18; ResNet-34; ResNet-50; ResNet-101 | ODIR2019 | Image; Medical | *Advantages:* This method has lower computational complexity<br>*Disadvantages:* The algorithm does not address the problem of class imbalance |
| ELIFT [32] | BR; ML-KNN; RAKEL; ECC; EPS; LIFT | Birds; Enron; Arts; Corel5k; Pascal | Audio; Text; Image | *Advantages:* The algorithm works better on high-dimensional data sets<br>*Disadvantages:* It did not consider the correlation between labels |
| ACkEL [48] | BR; CC; CLR; LIFT; ACkEld; ACkElo | Birds; CAL500; CHD49; Emotions; Flags; Foodtruck; Genbase; GpositiveGo; Medical; PlantGo; etc | Audio; Music; Text; Biology; Medical; Image; etc | *Advantages:* The algorithm solves the problem that the transformed classes are difficult to be separated in the feature space<br>*Disadvantages:* The algorithm did not find the best $K$ for different data sets, did not try other base classifiers, and the complexity of training was relatively high |
| AESAKNNS [33] | NB;HT;OBML;MLkNN; MLSAMPkNN; MLSAkNN; etc | Birds; Flags; Yelp; Stackex; Ohsumed; Water-qual | Biology; Text; Audio; Music; Pictures; Chemistry; | *Advantages:* The algorithm can be adapted to the concept drift problem while using a random subspace to increase the diversity of base classifiers<br>*Disadvantages:* It does not adapt to problems such as adaptive windows and additional hyperparameters, nor does it adapt to dynamic ensemble scales |

**Table 5** (continued)

| Algorithm | Comparison algorithms | Experimental datasets | Test fields | Advantages and disadvantages |
|---|---|---|---|---|
| MLSAkNN [31] | ML-kNN; MLkNNP; MLSAMPkNN; MLkNNPA; etc | Birds; Flags; Yelp; Stackex; Ohsumed; Genbase; etc | Biology; Text; Audio. Music; Pictures; etc | *Advantages:* It can quickly detect and respond to various types of conceptual drift, and none of the proposed mechanisms require adaptation<br>*Disadvantages:* It does not add an adaptive feature selection mechanism, nor does it introduce techniques to exploit label correlation |
| MLSAMPkNN [30] | ML-Knn; kNNP; kNNPA; SAMKNN; etc | Birds; Flags; Yelp; Stackex; Ohsumed; Genbase; etc | Biology; Text; Music; Pictures; Audio; etc | *Advantages:* It can react to various variations, drifts, imbalances and noise that may occur in the data stream<br>*Disadvantages:* It does not add a penalty mechanism that allows it to learn from a very sparse stream of labels |
| ML-BTC [36] | LIFT; RAKEL; MLKNN; G3P-KMELC; LSF-CC; MLTL-HOMER; etc | Flags; CAL500; Scene; Yeast; Enron; Image; Water Quality; Corel; CHD49; Delicious | Image; Text; Audio; Biology; Chemistry; | *Advantages:* It can solve the inherent problem of class imbalance in multi-label data<br>*Disadvantages:* The algorithm is more parameter-dependent |
| BNCC [39] | BR; CLR; CC; GCC; ECC | Bibtex; Enron; Scene; Yeast; Genbase; Mediamill; Emotions; etc | Text; Music; Biology; Audio; Images; etc | *Advantages:* While improving the performance of traditional CC, the algorithm can help determine better label sequences for CC<br>*Disadvantages:* The algorithm does not validate the effectiveness of different base classifiers and does not design an alternative scoring function |
| ML-KELM [26] | Rank-SVM; ML-KNN BoosTexter ELM | Emtions; Yeast; Scene; Core16k; etc | Music, Biology, Image, Text | *Advantages:* It requires fewer parameters to be tuned and performs better in large-scale data environments<br>*Disadvantages:* The algorithm does not analyze the sensitivity of the parameter settings and also does not use sparse coding to improve the classification performance |
| ML-ELM-RBF [47] | RELM; ML-ELM; etc | Arts; Education; Health; Entertainment; Science; etc | Text | *Advantages:* The algorithm obtains better results in terms of classification results and running time<br>*Disadvantages:* The algorithm could be further investigated for methods to determine the number of hidden layers and each neuron as well as methods to determine the parameters |

**Table 5** (continued)

| Algorithm | Comparison algorithms | Experimental datasets | Test fields | Advantages and disadvantages |
|---|---|---|---|---|
| ML-AP-RBF-Lap-ELM [25] | ML-RBF-RELM; ML-ELM-RBF; ML-RBF; RELM; ML-KNN | Yeast; 20NG; Scene | Biology, Pictures, Text | *Advantages:* It considers the structural relationship between low-dimensional data<br>*Disadvantages:* The accuracy and generalization of the algorithm still need to be improved while maintaining stability |
| ML-CK-ELM [27] | ML-ELM-RBF; RELM; ML-RBF; ML-KNN | Yeast; Scene; Art; etc | Biology, Text | *Advantages:* It eliminates the need for random parameter tuning, while computation time and memory storage are drastically reduced<br>*Disadvantages:* It does not examine the different ways in which the nuclei can be combined |

much information. The algorithm under tag association can obtain better accuracy than the algorithm under non-tag association. If the application has higher requirements for accuracy, it can be achieved by solving the related problems.

MLSAMPKNN, AESAKNNS and MLSAKNN can handle the concept drift that occurs in the data stream. Penalty mechanisms and enable/disable labels can be introduced when the algorithm needs to handle concept drift. If one wants to increase the diversity of the underlying classifier, measures of feature subspaces can be introduced. LdSM and ML-KELM can experiment on large-scale data. ML-KELM performs better and is more stable on large-scale data. For the dataset RCVLV2, the Hamming loss of ML-KELM is about 5.8% and the coverage is about 12.8%, while the Hamming loss of the comparison algorithm RANK-SVM is about 7.1% and the coverage is about 10.8%, and the smaller values of two evaluation metrics indicate better classification performance. And the accuracy of LDSM decreases compared with the comparison method. But it has lower complexity and shorter prediction time. If one wants to have better classification performance on large data sets while spending less time, one can introduce the kernel extremum learning machine principle or use a tree structure that facilitates balanced splitting to maintain a high degree of purity of the child nodes and has penalties for overgrowth.

## 2.2 Semi-supervised learning

In practical applications, obtaining fully labeled instances is expensive and time-consuming, and using incompletely labeled data for training is a practical approach. Let $D = D_L + D_U$ be a set of instances, where $D_L$ and $D_U$ are the sets of labeled and unlabeled instances, respectively. The task of semi-supervised MLC is to construct a classification function $f$: $D_L \cup D_U \rightarrow 2^L$. This section presents inductive and transductive methods, where inductive methods involves the optimization of the prediction model, while the transduction methods optimizes the prediction directly.

### 2.2.1 Inductive methods

Semi-supervised multi-label inductive methods typically extend SL algorithms to allow them to handle unlabeled data. This section provides an overview of the algorithms from three main perspectives: wrapper, clustering and others.

**2.2.1.1 Wrapper algorithms** Wrapper can be divided into co-training, self-training and boosting. Li et al. [50] fused the algorithms of MLkNN [51] and FESCOT [52] to form COMN algorithm. COMN is trained on the same dataset by using a pair of MLkNN classifiers with two different sets of parameters. Both classifiers label unlabeled instances

and provide each other with training datasets. SSR-CT [53] is a co-training method based on semi-supervised regression. During the co-training process of the algorithm, each learner first makes predictions for unlabeled instances, and then selects and adds the most confidently labeled unlabeled instances to another learner's training set to improve its performance. The algorithm iterates until the stopping condition is satisfied and the final prediction of the test data is the average of the two learners' predictions. Each base classifier in SSkC [54] is trained in a co-training fashion. To avoid the accompanying set giving biased labeled predictions, each accompanying base classifier is required to label only its accompanying instances. Once the algorithm updates all base classifiers, the labeling decision threshold is recalibrated to satisfy the target loss function and the importance of features is re-evaluated using both labeled and unlabeled instances.

SS-MLLSTSVM [55] is a semi-supervised multi-label least squares double SVM. It introduces the least squares idea into each subclassifier of MLTSVM, so that each subclassifier only needs to solve a linear system of equations, and introduces a manifold regularization term in each subclassifier, which can make full use of the geometric information in unlabeled and partially labeled samples. LP-MLTSVM [56] proposed a new two-stage classification method. In the first stage, the labels of the unlabeled training data are determined by using a smooth graph constructed by manifold regularization. In the second stage, a multi-label classifier is built.

Zhan et al. [57] used the under-inductive setting in their algorithm. In each round of co-training, the dichotomy of the feature space is learned by maximizing the diversity between the two classifiers induced on the dichotomous feature subset. CobMLkNN [58] extends the paradigm of co-training using the multi-label kNN algorithm. The principle is to identify the k-nearest instances of each test instance and calculate the number of neighbors belonging to the same label. It then uses the maximum a posteriori principle to determine the set of labels for each test instance.

Nowadays, many applications in life can generate more and faster data than ever before, but most co-training methods cannot deal with this problem. For this reason, Chu et al. [59] first used the sliding window mechanism to divide the data stream into data blocks and trained a basic classifier for each data block using COINS, and then an ensemble model with a WCOINS classifier is generated to adapt to the data stream environment containing a large amount of unlabeled data. At the same time, a new class emergence detection mechanism is introduced to detect the emergence of new classes in the data block. When a new label is detected, the classifier is retrained on the current data block and the integrated model is updated.

Self-training is another technique most commonly used in SSL [60]. Santos et al. [61] proposed two methods of applying semi-supervised technology of self-training, namely SSLP and SSRAkEL. These methods are based on their corresponding monitoring methods LP and RAkEL. Santos et al. [62] proposed a self-training method for hierarchical multi-label problems, HMC-SSLP and HMC-SSRAKEL. But these two methods are associated with the random selection of unlabeled instances for labeling. To solve this problem, Rodrigues et al. [63] proposed to use confidence parameters in the automatic label allocation process in combination with data stream features. First, the algorithm uses the labeled data stream set to train the classifier and calculates the confidence coefficients for all unlabeled samples in the dataset. Then, it sorts the unlabeled samples in descending order based on the standard deviation and selects the top n examples in the sort. Finally, a label is assigned to all selected examples and the newly labeled examples are moved to the labeled dataset.

MH [64] is a well-known extension of AdaBoost [65] in MLC, which efficiently handles multi-label problems by transforming MLC problem into several binary classification problems. Zhao et al. [66] also proposed a semi-supervised MLC algorithm based on AdaBoost, which proposes to use conditional variance as regularization to exploit information from unlabeled data and encourages it to find hypothetical labels for unlabeled data, which helps drive the algorithm to produce better combinatorial classifiers.

**2.2.1.2 Clustering algorithms** Clustering algorithms can divide instances into labeled and unlabeled sets, and then assign labels to unlabeled instances by classification. AHMED [67] uses fuzzy clustering, which allows each data point to belong to multiple clusters. First, the algorithm updates the dimensional weights and cluster membership values. After that, it updates the centroids of the clusters and updates the summary statistics. In this step, it determines $K$ nearest neighbor clusters for each test data point. This distance is computed in the subspace of the clusters. If $K$ is greater than 1, the algorithm calculates the probability of a class by multiplying the inverse of the distance between the representation of the class and the subspace, and then sums each class over all $K$ nearest clusters.

FS-MLSS-KSC [68] uses the kernel spectrum clustering algorithm as the core model and integrated information from labeled data points into the model through regularization terms. It then implements the propagation of multiple labeled data points to unlabeled data points by combining correlations between labels. The algorithm uses the Nystrom approximation to construct an explicit feature map and solves the optimization problem in the original function. $\text{OPFSEMI}_{\text{mst+knn}}$ [69] uses the optimal path forest framework. Since misclassified samples usually appear

at the boundary between clusters, this method reduces the error in label propagation in the training set by re-propagating labels from the maximum of probability density function. In addition, the algorithm gives higher priority to training samples closer to their maxima and assigns their labels to new samples during classification.

Pham et al. [70] used a greedy method to select class-label specific features as an extension of the LIFT algorithm, as well as a label-free data consumption mechanism from text classification using a semi-supervised clustering algorithm. In the clustering phase, the algorithm uses clustering to identify components in labeled and unlabeled instances based on the highlighted labels. In the classification phase, it determines the nearest instance clusters and assigns labels to the unseen instances. Ha et al. [71] proposed TESC. In the clustering phase, TESC uses the labeled text to capture the silhouettes of the text clusters. Next, it adds unlabeled texts to the corresponding clusters to adjust their center point. In the classification phase, it uses kNN to find the most recent clusters and returns the label set of the found clusters as the label set of the new data instance. MCUL [72] uses clustering-based regularization terms to discover unobserved labels in the dataset and uses the specific label features learned to describe their semantics and use label correlation to overcome the problem of missing labels.

**2.2.1.3 Other algorithms** In addition to wrapper and clustering, inductive methods also use kernel norm or low-rank regularization. SLRM [73] uses kernel constant regularization on maps to efficiently capture label correlations and introduces stream regularization to capture the internal structure between data. In the regularization, when two instances are close in the feature space, their new representation based on the map should be close. In this case, the mapping is able to capture the intrinsic geometric structure between instances in the feature space and label space. Sheng et al. [74] propose an adaptive low-level SSL multi-label algorithm. In this algorithm, the intermediate feature space for learning labeled and unlabeled training samples is reduced by a low-rank matrix, and the multi-label classifier is trained by an adaptive SSL strategy.

In order to solve the noise problem in the examples, SUN et al. [75] proposed robust semi-supervised multi-label learning based on three-low rank regularization. The algorithm first introduces a linear self-representative model, which uses label correlation to recover the matrix of labels that may be noisy. Then, it uses low-rank representation to construct a low-rank polarity matrix to capture the global relationship between labeled samples and unlabeled samples. The graph Laplacian regularization is constructed by using the pair similarity matrix defined above to obtain information on the geometric structure of the labeled and unlabeled samples. The prediction models of different labels are connected in series into a matrix and the matrix tracking norm is introduced to capture the correlation and complexity of the control model. CORALS [76] optimizes all possible labels by minimizing cost-sensitive ranking losses, using dual low-rank regularization to capture the corresponding correlations and using sparse regularization terms to constrain the sparsity of noisy information.

### 2.2.2 Transductive methods

Transductive methods in SSL are graph-based, either explicitly graph-based or implicitly graph-based [12]. This section mainly explains graph-based construction and graph-based weighting of transductive methods.

**2.2.2.1 Graph-based construction** Zha et al. [77] proposed a graph-based SSL framework, which can simultaneously explore the correlation between multiple labels and label consistency on the graph. Specifically, the framework employs two types of regularizers. One is used to select the label smoothing on the graph, and the other is used to address that the multi-label assignment of each example should be consistent with the inherent label correlation.

In some classification tasks, local feature descriptor-based methods are more robust to intra-class variation than global feature-based methods [78]. LSS [78] outperforms the global feature-based GRF algorithm in some classes. The performance of LSS depends on the accuracy of the feature matching context. Bao et al. [79] proposed a semi-supervised multi-label image labeling algorithm, which based the propagation of labels on virtual local label representation rather than on the whole image representation, and proposed an effective multiplication iterative method to optimize the objective function. Later, Jiang et al. [80] proposed an extended algorithm of graph learning based on local and global consistency, named Multi-label Dependent semi-supervised learning (MCSL). It incorporates the intrinsic correlations between functional classes into protein function prediction by utilizing the relationships provided by PPI network and functional class network. The classification function should be smooth enough on the subgraph where the respective topologies of the two networks are well matched.

The complexity of data distribution in practical applications makes it difficult for the algorithm to choose the appropriate parameters. To address this problem, Liu et al. [81] proposed an SSL framework for MLC based on kernel norms. The framework uses kernel normalization for class-level smoothing, uses criterion functions to construct class graphs adaptively, and introduces a non-greedy iterative algorithm to solve the criterion functions. It also proposes two algorithms based on the kernel norm. Formula 2

is NML-GRF, and Formula 3 is NML-LGC. $\boldsymbol{F}$ is the prediction label matrix, $\boldsymbol{Y}$ is the label matrix, $\boldsymbol{M}$ is $(FF^T)^{-\frac{1}{2}}\boldsymbol{L_n}$ is the normalized Laplacian matrix of the instance graph, $L_n = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.

$$\min \mathrm{tr}\left((\boldsymbol{F}-\boldsymbol{Y})^T \mathrm{U}(\boldsymbol{F}-\boldsymbol{Y})\right) + \mathrm{tr}\left(\boldsymbol{F^T L F}\right) + \gamma \mathrm{tr}\left(\boldsymbol{F M F^T}\right) \quad (2)$$

$$\min \mathrm{tr}\left((\boldsymbol{F}-\boldsymbol{Y})^T \mathrm{U}(\boldsymbol{F}-\boldsymbol{Y})\right) + \mathrm{tr}\left(\boldsymbol{F^T L_n F}\right) + \gamma \mathrm{tr}\left(\boldsymbol{F M F^T}\right) \quad (3)$$

Ghosh et al. [82] proposed two different graph-based methods, namely Label Correlation Propagation-GRF (CP-GRF) and Weighted Label Correlation Propagation-GRF (WCP-GRF). CP-GRF involves propagation on a label correlation graph for each instance. WCP-GRF is an extension of the CP-GRF method, its correlations are not only propagated from other related labels, but are also based on proximity to a specific example. SMILE [83] used the known labels and supplementary labels of labeled instances and unlabeled instances to train a graph-based semi-supervised linear classifier and directly predicts the labels of new instances that are completely unlabeled. Behpour et al. [84] developed adversarial Robust Cuts (ARC), using learning tasks as a minimax game between predictors and "label approximators" based on minimal cost graph cuts. ML-GCN [85] uses GCN to embed node features and graph topology information. The algorithm randomly generates a label matrix with the same dimensionality of the label vector as the node vector before the last convolution operation. During training, it concatenates the label vector and the node vector as inputs to a relaxed jump graph model to detect node-label correlations and label correlations.

In images, many algorithms are proposed for annotation functions. AHL [86] is a multi-label image labeling method based on adaptive hypergraph learning. The algorithm preserves the local geometric structure of the data in a high-order manner and obtains a potential feature space by adding feature projections in which multiple labels can be efficiently and robustly assigned to unlabeled instances. WeSed [87] uses weakly weighted pairwise ranking loss for weakly labeled images and triple similarity loss for unlabeled images.

Wang et al. [88] proposed a dual low-rank regularized multi-label learning model. The algorithm introduces a dual-trace regularization to capture the correlation between different label prediction models in the feature space and a linear self-recovered model to recover the noisy training label matrix in the learning phase. MLRMG [89] creates multiple graphs based on a randomly selected subset of features, learns the labeling function on each graph by optimizing a semi-supervised loss function, and finally, it votes on multiple graphs to determine predictive labels for unlabeled data. Song et al. [90] introduced the idea of label embedding to capture the network topology and higher-order multi-label correlations. The similarity of label embedding and node embedding can be used as a confidence vector to guide the label smoothing process, forming a marginal ranking optimization problem to learn the second-order relationships between labels.

Boulbazine et al. [91] proposed an online semi-supervised multi-label classifier based on the Growing Neural Gas (GNG) algorithm. The main principle of the algorithm is to associate a prototype consisting of two vectors with each neuron $k$. These two vectors are updated for each neuron during the learning process and used for prediction of unknown label vectors. Li et al. [92] extended the graph-based SSML to MLC, and also investigated three graph regularization methods: Gaussian Field and Harmonic Function (GFHF), Local and Global Consistency (LGC), and Manifold Regularization Modification (MR), and propose a semi-supervised multi-label decomposition framework for the NIALM problem. MGLP [93] uses multi-level neighborhood information granularity and a three-way decision method, where the three-way decision method can be used to select unlabeled data for further annotation. Through the iterative process of label propagation, data annotation and data subset update, the final pseudo label accuracy of unlabeled data is improved.

**2.2.2.2 Graph-based weighting** The classic label propagation algorithm gives a finite weighted graph $G=(V, E, W)$, where $V$ is composed of the dataset $\boldsymbol{X}=\{x_i, i=1, \ldots, n\}$ and $E$ is composed of $V \times V$, $W$ is a non-negative symmetric weight function, and the algorithm interprets the weight $w(i, j)$ as a similarity measure between vertices $x_i$ and $x_j$. If $\rho$ is a distance metric defined on the graph, then the similarity matrix can be constructed as follows:

$$W(i,j) = h\left(\frac{\rho\left(x_i, x_j\right)^2}{\mu\sigma^2}\right) \quad (4)$$

One disadvantage of label propagation is that it does not handle multi-class or MLC problems well due to the lack of interaction between labels in different classes. Reference [94] proposed a dynamic version of label propagation. For the dynamic propagation algorithm, it sets the similarity between non-adjacent points to 0, i.e. assumes that local similarity is more reliable than distant ones. Local similarities can be propagated to non-local points through a diffusion process on the graph. At the same time, KNN is used to test the local distance. The similarity matrix is constructed as follows:

$$Wi,j = \begin{cases} W(i,j), if\left(x_j \in KNN\left(x_i\right)\right) \\ 0 \end{cases} \quad (5)$$

Dharmadhikari et al. [95] used the KNN method to reweight adjacency matrix $\boldsymbol{A}$, and uses a cosine similarity measure to represent edge weights and generate a matrix $\boldsymbol{W}$ through a graph segmentation process. Such graph specification can improve the efficiency of the label inference stage. Lucena et al. [96] extended MLkNN algorithm to SSL. The algorithm creates weight matrices and diagonal matrices using instances of the partially labeled dataset. The formula for the weight is as follows:

$$W_{ij} = e^{-\frac{x_i - x_j}{2\delta^2}} \tag{6}$$

Among them, they transformed the training dataset into graph $\boldsymbol{G}$ ($\boldsymbol{V}$, $\boldsymbol{E}$), $e \in \boldsymbol{E}$. $\boldsymbol{W}_{ij}$ defines the similarity between nodes $i$ and $j$.

Gang et al. [97] proposed to construct two graphs at instance level and category level, respectively. For instance-level, the definition of the graph is based on labeled and unlabeled instances, where each node represents an instance and the weight of each edge reflects the similarity between the corresponding paired instances. For the class hierarchy, a graph is constructed based on all classes, where each node represents a class, and the weight of each edge reflects the similarity between the corresponding pairwise classes.

To make the algorithm more robust to noise and incomplete image labels, Cevikalp et al. [98] argue that it is important to use a robust ramp loss. The algorithm passes the labels of the labeled data samples to the nearest unlabeled samples and uses the similarity score to control the reliability of the label assignment. The weight formula of the algorithm is as follows:

$$\begin{aligned} &\min_{w} \frac{\lambda}{2} trace\left(\boldsymbol{W}^T \boldsymbol{W}\right) + \sum_{i=1}^{l+u} \sum_{j=1}^{C_{x_i}^+} \sum_{k=1}^{C_{x_i}^-} s_i L\left(r_j\right) R_s \\ &\left(\boldsymbol{W}_j^T x_i - \boldsymbol{W}_k^T x_i\right) + K \sum_{i=1}^{l+u} \sum_{j=1}^{m} s_i R_s \left(y_{ij} \left(\boldsymbol{W}_j^T x_i\right)\right) \end{aligned} \tag{7}$$

where $u$ is the unlabeled label, $x_i$ feature vector, $C_{x_i}^+$ and $C_{x_i}^-$ represent the positive and negative labels of $x_i$. $s_i$ is equal to 1, $i = 1, \ldots, l$, $r_j$ is the rank. $L(.)$ is the weighting function for different levels and $\boldsymbol{W}$ is the weight matrix. $\lambda$ is the regularization parameter and $K$ is a user-defined parameter that controls the slope loss weight.

### 2.2.3 Summary

This chapter provides a tabular overview of the time complexity, application areas, and advantages and disadvantages of individual algorithms.

**2.2.3.1 Time complexity** Time complexity refers to the amount of computational effort required to execute an algorithm. It can measure the efficiency of the algorithm, pointing out the relationship between the computational workload performed by the algorithm to solve the problem and the size of the problem. The time complexity of the individual algorithms involved in this paper is shown in Table 3. We can find that it is mostly related to the number of instances and feature dimensions. Individual algorithms are also related to the number of iterations performed, and the size of the model.

SMILE runs faster than all comparison algorithms with 30% missing labels. On the three datasets, SMILE only takes a total time of 280.51 s, while the comparison algorithm MLML consumes 7615.33 s. SS-MLLSTSVM, although it needs to compute the Laplace matrix for the whole sample, still has a faster running speed. On Flags, it takes only 0.037 s, while the comparison algorithm BPMLL takes 4.241 s. The running speed of CORALS decreases significantly with the increase in the number of instances, label classes and feature dimensions. This is because the method focuses on checking the correctness of each class label (Table 6).

**2.2.3.2 Analysis of performance of algorithms** The test domains for semi-supervised multi-label learning are generally text, audio, images, biology, and music. Among them, the algorithms focusing on text classification include algorithms SISC [67], GB-MLTC [95], MULTICS [70]. SISC can determine clusters in subspaces of high-dimensional sparse data. GB-MLTC can use cosine similarity measures that may ignore certain aspects of semantic relationships between text documents that may affect accuracy. MULTICS can be derived from the text classification of the unlabelled data consumption mechanism.

In the field of image, LSS [78], SSML [92], WeSed [87] and AHL [86] algorithms are suitable for multi-label image labeling, and LSS can obtain better results when matching more images. SSR-CT [53] uses regression and cooperative algorithms to classify and predict images, but it is easily affected by noise. But WeSed does well with noisy data. CNN + RMLC [98] can remove error samples well to expand the training set and suitable for retrieval of large-scale images.

In the face of large amounts of data, multi-label data stream algorithms are particularly important. Both cooperative training and self-training algorithms in inductive methods can reasonably process data stream data, such as the algorithm Rodrigues [63].

Finally, we summarize semi-supervised classification algorithms from the perspective of label non-correlation and label correlation through Fig. 2, which includes test fields, advantages and disadvantages.

## 3 Application field

MLC problems have attracted more and more researchers' attention due to their wide application [13]. In the next few sections, MLC algorithms will be described from the fields of image classification, text classification, and others.

### 3.1 Image field

Image classification is a difficult task that has attracted great attention from the research community recently. Image classification is more suitable to use MLC algorithms for classification. This is because most images can be described with multiple labels to describe their semantic content, such as objects, scenes, actions, attributes, etc. [41]. This section is mainly introduced from the medical field and the remote sensing image field.

#### 3.1.1 Medical field

Image of MLC has a wide range of applications in the medical field, such as chest X-rays, electrocardiograms, brain CT, eye diseases, etc. The algorithms of MLC can make up for the shortage of doctors and reduce the workload of doctors.

Chen et al. [42] proposed a novel label co-occurrence learning model for multi-label chest X-ray image classification, which explores potential co-occurrence labels in images by using label co-occurrence and dependent information. Guan et al. [99] proposed the CRAL model to solve the problem of multi-label chest disease classification on chest X-ray images. It predicts the presence of multiple lesions in a particular category of attentional view and suppresses disorders in unrelated categories by assigning smaller weights to the corresponding features. Chougrad et al. [100] used SGD with exponentially decaying learning rate to effectively improve domain adaptation so that the model can maximize learning over new domains for better classification prediction of mammograms.

Cai et al. [102] proposed a method for arrhythmia based on electrocardiogram data set, which can detect 55 kinds of heart disease symptoms at the same time, and call it Multi-ECGNet. This model proposes a complete set of ECG monitoring analysis, modeling methods and research ideas of an end-to-end deep learning model, and at the same time is superior to ordinary cardiologists in terms of indicators. Li et al. [43] proposed a multi-label slice-dependent learning model called SDLM. It is a sequence-to-sequence model that effectively learns image features and slices dependencies in an end-to-end manner. He et al. [44] proposed a model that considers patient-level diagnosis and multi-label disease classification that are associated with binocular eyes. Three models are proposed. The first is the CNN model, which can classify patient-level multi-label eye diseases, and can handle seven eye diseases at the same time through a single network, and the second is a novel module SCM, which is designed to effectively integrate from Control the function of CFP extraction. Ou et al. [103] proposed bilateral feature Enhancement Network, which uses the interaction between bilateral fundus images to enhance the extracted feature information. Feature information from images with different resolutions extracted by extended convolution is superimposed, enriching the feature images and thus capturing more disease features.

Xu et al. [104] explored easily accessible labels to help classify lesion types, used the label of lesion type and patient ID to construct a loss function based on DML and also used five-fold input to build a deep model using transfer learning. Finally, a five-fold mining algorithm for label selection training samples is proposed.

**Table 6** Table of time complexity analysis

| Algorithm | Dataset | Time complexity |
|---|---|---|
| SMILE [83] | Cal500; Bibtex; Delicious | $N^2C + N^2D + ND^2 + D^3$, $C$ is the number of distinct labels of the instance, $N$ is the number of instances, and $D$ is the number of features |
| AHL [86] | CUB; SUN; AWA; Corel5K; IAPR-TC12; ESP Game | $n^3 + d^3$, $d$ is the feature dimension and $n$ is the number of samples |
| SS-MLLSTSVM [55] | Flags; Emotions; Birds; Scene; Yeast | Linear: $O(n^2 log(n) + Kd^3)$, nonlinear: $O(n^2 log(n) + Kn^3)$, $K$ is the number of labels, $n$ is the number of instances, and d is the distance |
| CORALS [76] | Emotions; CAL500; Genbase; Medical; Corel5k; Pascal07; Delicious; ESPGame | $O(t \times (q^3 + r^3))$, $t$ is the iteration time to update the model, $q$ is the class label, $r$ is min($d$,$q$) |
| MCUL [72] | Bibtex; Corel16k001; Medical; Stackex | The time complexity of $\|S - HH^T\|_F^2$ in the objective function is $O\left(n^2(d + l + n)\right)$ |
| MGLP [93] | Wine; Lonosphere; Breast; Heart; Yeast; Image; Wireless; QSAR | The worst case is $O(kun^2)$, $n$ is the number of instances, $u$ is the unlabeled instances, and $k$ is related to KNN |

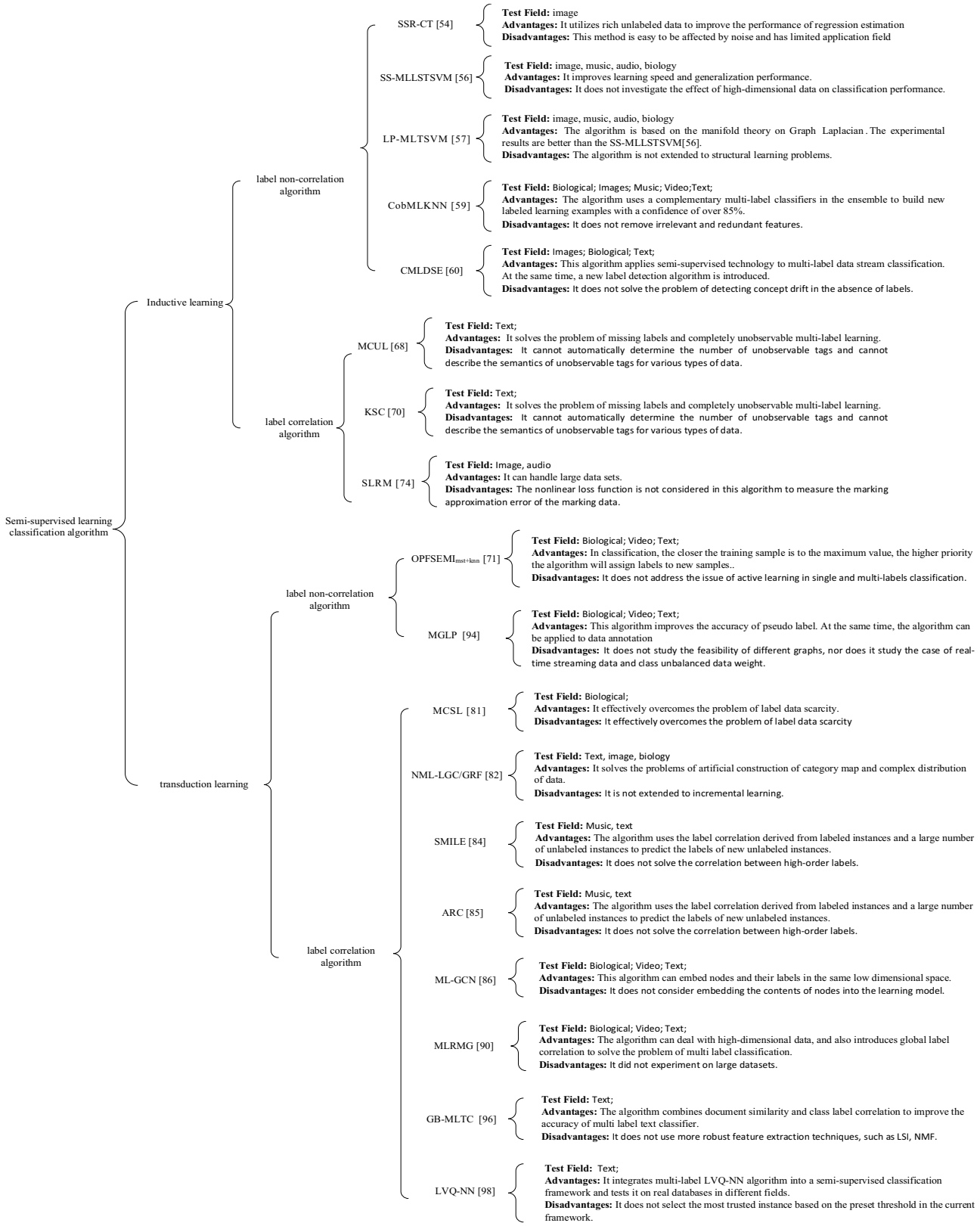


**Fig. 2** Summary of SSL algorithms

#### 3.1.2 Remote sensing image

Multi-label image classification plays an important role in the complex content of remote sensing images, and has triggered some related studies in the past few years [105].

Alshehri et al. [106] proposed a deep learning model based on a codec NN architecture with a channel and spatial attention mechanism to deal with remote sensing-based drone problems. The model is based on the task of a pre-trained CNN encoder module and converts the input image into a set of feature maps using an appropriate combination of features. The task of the decoder module based on LSTM network is to generate the classes present in the image in a sequential manner. Hua et al. [105] proposed an attention-aware labeled relational inference network based on remote sensing technology. The network consists of three basic modules. The learning module of labeling plots by label aims to extract high-level elements specific to labels; the attention area extraction module generates attention label-specific functions; the label relationship inference module uses the output derived from the previous module the label relationship is used to predict the existence of the label.

A novel multi-attention drive system was proposed by Sumbul et al. [24] in 2020. The system is mainly divided into four modules. The first module extracts preliminary local descriptors of remote sensing image bands that can be associated with different spatial resolutions. The second module is implemented by a two-way RNN architecture, in which LSTM nodes enrich local descriptors by considering the spatial relationship of local regions. The third module is implemented through a patch-based multi-attention mechanism, which takes into account the co-occurrence of multiple land cover categories. The last module uses these descriptors to classify multi-label remote sensing images. Chaudhuri et al. [107] proposed model based on four main steps: The first step is to segment each image in the archive and extract the features of each region. The second step is to construct the image neighborhood map and use the relevant label propagation algorithm. The third step uses a novel region labeling strategy to associate the class label with the image region, and the last step uses a sub-image matching strategy to retrieve images similar to the given query image.

Dai et al. [108] proposed a new CBIR model. Combining spatial and spectral descriptors, this model achieves image retrieval through a novel remote sensing image retrieval method based on sparse reconstruction, considers a new label likelihood metric, and extends the original sparse classifier to single-label and multi-label remote sensing image retrieval problems, proposing a strategy to exploit the sensitivity of the sparse reconstruction-based approach to different dictionary words. Koda et al. [41] proposed an SVM-based MLC method to achieve accurate land cover classification of remote sensing images. The model that enhances the smoothness of the entire image is called Spatial Structured SVM (SSSVM).

### 3.2 Text classification

The application field of text classification can also be solved by using MLC algorithms. The main application fields are sentiment classification and medical biology classification.

#### 3.2.1 Sentiment classification

Multi-label sentiment classification is a subtask of text sentiment classification. Its purpose is to identify coexisting emotions (such as joy, anger, anxiety, etc.) expressed in the text. Due to its broad potential, it has attracted the attention of researchers [64].

He et al. [109] proposed a JBNN, which can effectively solve the problem that binary networks ignore the correlation between labels. In JBNN, the representation of text is replaced by a set of logistic functions instead of softmax function, and multiple binary classifications are performed simultaneously in a single neural network framework. In addition, the relationships between labels are obtained by training a joint binary cross-entropy loss. Yu et al. [110] proposed a new transfer learning architecture. The model uses a shared LSTM layer to extract shared emotion features for emotion and sentiment classification tasks, and uses a target-specific LSTM layer to extract specific emotion features that are only sensitive to people's emotion classification tasks. Fei et al. [111] proposed a TECap, which can learn potential topic information without external knowledge, thereby promoting multi-label sentiment classification.

Alzu'Bi et al[112] proposed a model to solve the sentiment analysis of Arab social media. To make the annotated data set more accurate, the model uses a mediation process to check and update the annotated data set. Bravo-Marquez et al. [113] proposed a model of annotated sentiment dictionary. The model combines word-level functions and learning techniques to efficiently accomplish this task, and can use unlabeled tweets to identify emotional words from any collection of specific fields. Kim et al. [114] proposed an attention-based classifier. The model consists of an attention mechanism and multiple independent CNN, and its performance is further improved through preprocessing of emoticons and the use of additional dictionaries.

Mulki et al. [115] developed a Tw-StAR to identify emotions embedded in Arabic, English and Spanish tweets. The model performs one or more combinations of preprocessing techniques on the tweets, adopts the BR conversion strategy, and uses the TF-IDF scheme to generate tweets. Alhuzali et al. [116] proposed a SpanEmo model, which

uses multi-label sentiment classification for span prediction, learns the association between labels and words in sentences, and introduces a loss function. Hyun et al. [117] proposed a deep learning-based model combining linguistic embedding and sentiment embedding for text classification in a CL-AFF shared task, and sentence features extracted from the embedding model were used as a TextCNN to provide input for text classification. Ying et al. [118] chose the popular BERT language model to provide general language knowledge for modeling sentences. They used a twitter-specific preprocessor to decode twitter-related expressions, introducing a two-step training process to integrate common sense and detected domain knowledge for sentiment classification. Ameer et al. [119] proposed a large benchmark corpus for multi-label emotion classification tasks, which uses content-based methods, in-depth learning and transfer based learning methods to classify the corpus at the same time.

#### 3.2.2 Medical biology classification

Multi-label text classification plays an important role in the field of information retrieval and has had an impact on information retrieval in the field of medical biology. [120].

Du et al. [121] proposed the ML-Net model, which is a novel end-to-end deep learning framework. The model is an efficient and scalable method that combines the label prediction network with an automatic label number prediction mechanism, and it does so by using the prediction confidence score for each tag and deep contextual information in the target document. Glinka et al. [122] proposed a model to improve the feature selection method of multi-label medical text classification, investigating filter and wrapper methods and hybrid methods. Hughes et al. [123] allow automatic generation of context-based, rich representations of health-related information. They extracted urgent semantics from a corpus of medical texts and classified text fragments at the sentence level using CNN. Yogarajan et al. [124] use multi-label variants of medical text classification to enhance the prediction of concurrent medical codes. A new embedding on health-related text compares several variants of the embedding model when dealing with the unbalanced multi-label medical text classification problem.

Wasimp et al. [125] proposed a classification model for multi-label questions for fact-based and list-based question processes for biomedical question answering systems. In the prediction stage, the list-type problems use the COPY LAT prediction model, and the fact-type problems use the BR LAT prediction model. Baumel et al. [120] proposed a HA-GRU. The model can use attention weights to better understand which sentences have the most impact on decision-making and which words in the sentence have the most impact on each decision. At the same time, it can find the sentence with the highest score in each label and pass this most important find the word with the highest score in the sentence. RBA [126] is a rule-based algorithm developed using the dictionary method. It uses labels to train attention-directed RNNs to classify reports as positive reports for one or more diseases or normal reports for each organ system.

#### 3.2.3 Other fields

In addition to image and text classification, the MLC method has been widely used in other aspects. This chapter will introduce the application of MLC from two aspects of music and video.

Oramas et al. [127] proposed a multi-label music genre classification model using deep learning architecture. The model combines learning-based feature embedding with the latest deep learning methods. For each album, it collects cover images, textual comments, and audio tracks. Zhao et al. [128] proposed a model to classify multi-label music styles through user comments. The model is divided into two mechanisms, a label graph-based neural network mechanism responsible for classifying music styles based on the correlation between comments and styles, and a soft training-based mechanism introducing a loss function with a continuous label representation. Ma et al. [129] proposed a novel knowledge relation Framework, which uses graph CNN to automatically learn deep associations between styles. The approach focuses on integrating external knowledge and statistical information about musical styles to derive correct and complete dependencies between styles, alleviating the problems of overfitting and underfitting.

Kim et al. [130] proposed a NN method. This method uses an attention mechanism for space and time dimensions to ignore noisy and meaningless frames. The correlation between labels is considered by decomposing the joint probability of labels into condition items. Karagoz et al. [131] proposed an auto-encoder for reducing the dimensionality of video datasets, and combined the features extracted by the multi-objective evolutionary non-dominated sorting genetic algorithm and auto-encoder. Araujop et al. [132] proposed a video classification model based on the most advanced network architecture based on the intersection of linear algebra and deep learning. The layer in the classic form is denoted as “dense”, and the layer denoted by loops and diagonal lines are referred to as “compact”. The required size is represented by cascading and slicing. Jiang et al. [133] proposed a new system to achieve real-time and MLC of short videos. The system adds an activation adjustment layer before the output S-function to enhance the CNN's discriminative power for each label and uses label imbalance-aware training loss to reduce the effect of mostly irrelevant labels. Wu et al. [134] proposed a spatiotemporal location transformation framework for multi-label video classification tasks.

The framework uses the method of global action label co-occurrence and proposes a plug-and-play spatiotemporal label dependency (STLD) layer. STLD not only dynamically models tag co-occurrence in video through self-attention mechanism, but also completely captures spatiotemporal tag dependencies through cross attention strategy.

### 3.3 Summary

This section summarizes the MLC algorithm from the image field, text classification field and other application fields. In order to conveniently analyze the performance and advantages and disadvantages of the model, Table 7 summarizes the models mentioned in the paper from the aspects of algorithms, application fields, experimental data sets, advantages and disadvantages.

## 4 Evaluation metrics and public data sets

### 4.1 Evaluation metrics

It is important to choose the appropriate method to evaluate the performance of classification algorithms. In single-label learning, classification is considered as if the observations are correctly classified or unclassified, while in multi-label learning, classification can be considered as partially correct or partially incorrect [50].

Several metrics have been proposed to evaluate the performance of MLC algorithms. The most commonly used are one-error, accuracy, hamming loss, recall, rank loss, coverage, subset accuracy, average accuracy, and micro-F1. Specially, the subset accuracy is more than strict for the evaluation, it will result in very low metric values. The following is a detailed introduction to the evaluation indicators in MLC.

Given a multi-label dataset $S=\{(x_i, Y_i)\}_{i=1}^{n}$, where $Y_i$ is the true label of dataset $x_i$, $n$ is the number of instances in the dataset, $h(x_i)$ is the multi-label classifier, $I[\bullet]$ is the indicator function.

one-error: It evaluates the percentage of instances where the top-ranking labels are not in the relevant label set.

$$One-error=\frac{1}{n}\sum_{i=1}^{n} I\left[\min_{y_j\in Y_i} R_i\left(y_j \notin Y_i\right)\right] \tag{8}$$

Accuracy: It measures the fraction of correctly classified labels.

$$Accuracy=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|Y_i\cap h(x_i)|}{|Y_i\cup h(x_i)|}\right) \tag{9}$$

Hamming loss: It evaluates the frequency of misclassification of an instance label pair, that is, the instance predicts an irrelevant label or the relevant label is missed.

$$Hamming\ Loss=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}\sum_{j=1}^{l} I\left[h\left(x_i\right)_j\neq Y_{ij}\right] \tag{10}$$

Recall: It measures the average proportion of related labels for instances predicted to be related.

$$Recall=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i\cap h(x_i)\right|}{\left|Y_i\right|} \tag{11}$$

Rank loss: It calculates the score for incorrectly sorted label pairs.

$$Rank\ loss=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|\left(y_a,y_b\right):R_i\left(y_a\right)\rangle R_i\left(y_b\right),\left(y_a,y_b\right)\in Y_i^l\times\hat{Y}_i^l\right|}{\left|Y_i^l\right|\left|\hat{Y}_i^l\right|} \tag{12}$$

where $(y_a, y_b)$ is the pair class label for instance $x_i$ and $\hat{Y}=Y/Y_i$.

Coverage: It is an indicator used to averagely calculate the number of steps required to cover all relevant labels of an instance.

$$Coverage=\frac{1}{n}\sum_{i=1}^{n}\max_{y\in Y}R_i(y)-1 \tag{13}$$

Subset accuracy: Subset accuracy can evaluate all correctly classified instances of the label.

$$Subset\ accuracy=\frac{1}{n}\sum_{i=1}^{n} I\left[h\left(x_i\right)=Y_i\right)] \tag{14}$$

Average accuracy: The average accuracy is the average proportion of related labels that rank higher than a specific label.

$$Average\ precision=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\left|Y_i^l\right|}\sum_{y_i\in Y_i^l}\frac{\left|\left\{y_q\in Y_i^l:R_i\left(y_q\le R_i\left(y_i\right)\right)\right\}\right|}{R_i\left(y_i\right)} \tag{15}$$

where $R_i(y_i)$ is the predicted rank of the class label $y_i$ for an instance $x_i$.

Micro-F1: Consider the problem of class imbalance. It uses the F1 metric to evaluate each label separately and averages all labels.

$$Micro-F1=\frac{2\sum_{j=1}^{l}\sum_{i=1}^{n}y_{ij}h\left(x_i,y_j\right)}{\sum_{j=1}^{l}\sum_{i=1}^{n}y_{ij}+\sum_{j=1}^{l}\sum_{i=1}^{n}h\left(x_i,y_j\right)} \tag{16}$$

**Table 7** Application field algorithms

| Model | Application field | Experimental datasets | Advantages | Disadvantages |
|---|---|---|---|---|
| [42] | Medical image field | ChestX-ray14 | It proposes a new label co-occurrence learning method to exploit the correlation between pathologies | The imbalance of the label data set is not considered |
| [100] | Medical image field | datasets of arrhythmia-related to heart disease | It can obtain a good level in the field of cardiac arrhythmias | The algorithm did not attempt a better network model available |
| [43] | Medical image field | CQ500 dataset | The algorithm has very promising applications in CT images of the brain | It does not use image enhancement techniques on continuous images |
| [44] | Medical image field | ODIR2019 | The algorithm has low time complexity | It does not consider the correlation between labels and does not explicitly address the class imbalance |
| [99] | Medical image field | Chest X-ray14 dataset | The category residual attention mechanism is very effective for classifying chest X-ray images | The algorithm does not address the problem of unlabeled labels |
| [104] | Medical image field | Chest X-ray14 | It has a significant effect on improving the classification performance of large sample categories | This approach does not allow for a reasonable or adaptive selection of optimal parameters |
| [24] | Remote sensing image field | BigEarthNet benchmark file | The method achieves a reduced risk of overfitting by reducing the complexity of the model | The method does not apply a local region adaptive definition strategy based on the semantic content of remote sensing images |
| [105] | Remote sensing image field | UCM multi-label data set; DFC15 multi-label data set | This method considers the dependencies between the labels | The model is not extended to areas such as weakly supervised object detection and semantic segmentation |
| [107] | Remote sensing image field | UCMERCED file | By special processing, this method can be used for any kind of remote sensing images | Its accuracy is more sensitive to the choice of segmentation algorithm and region features |
| [108] | Remote sensing image field | UC Merced Land Use File | The system highlights the more informative features of the image to enhance the image representation | For large-scale operations, the system may require more retrieval time |
| [41] | Remote sensing image field | UAV data set; airborne image dataset | This method can model spatial continuity and output structure | The method has a high computational cost |
| [109] | Emotion analysis | Ren-CECps Corpus | The model obtained better results in terms of classification performance and computational efficiency | The model has only experimented on the Ren-CECps corpus |
| [110] | Emotion analysis | SemEval-18; SemEval-16; etc | The model captures the correlation between emotions through a labeled relational prior | It did not consider the correlation label imbalance issues |
| [112] | Emotion analysis | Twitter dataset | The model is used for sentiment classification of Arabic tweets | The model does not use a dictionary-based unsupervised approach |
| [114] | Emotion analysis | SemEval-2018 | The model preprocesses emojis and uses additional vocabulary to improve classification performance | The correlation between labels is not considered |
| [117] | Emotion analysis | OffMyChest dataset | The model uses emotional gloves to improve the classification performance of the imbalance class | The model does not use a valid way to convey the context of a post |

**Table 7** (continued)

| Model | Application field | Experimental datasets | Advantages | Disadvantages |
|---|---|---|---|---|
| [121] | Medical biology | Biomedical literature and clinical notes | It can accurately represent the context and dynamically estimate the number of labels | The model does not map the text to a high-dimensional representation of the document encoding network for further improvement |
| [123] | Medical biology | PubMed; Merck Manual; | Multi-layer convolutional deep networks can generate more optimal features in the training phase | No larger scale and finer-grained clinical classification were used to implement the proposed technology |
| [127] | Other music | MuMu dataset | This paper investigates the MLC of music genres from different perspectives and uses different data models | The ResNet in this paper cannot learn the underlying factors from the images |
| [128] | Other music | Dataset collected by Chinese music websites | This work is the first to explore a multi-label music style classification driven by reviews | It performs poorly in the low-frequency style in the training set |
| [129] | Other music | Douban Music; Amazon Music; | The model shows high efficiency in alleviating the underfitting problem | The model does not explore more efficient label representation methods to strengthen label dependencies |
| [131] | Other videos | MIR-Flickr; wireless multimedia sensor data set; | The model reduces the number of features and increases the Hamming score through a multi-objective optimization process | The model does not use other feature selection techniques to develop different types of autoencoders and multi-objective feature selection algorithms |
| [132] | Other videos | YouTube-8 M | The model allows for a trade-off between model size and accuracy | The correlation between labels is not considered |
| [133] | Other videos | Dataset provided in AI Challenger 2018 | The model achieved 86% accuracy on both the test set and 89.2% accuracy on its open validation set with a speed of 28 ms/sec | The semantic levels and motion vectors between different categories are not used to better handle labels from different conceptual levels |

**Table 8** Evaluation metrics of SL algorithms

| Algorithm | Evaluation metrics | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HL | SA | Acc | Pre | Re | F1 | Micro Pre | Macro Pre | Micro Re | Macro Re | Micro F1 | Macro F1 | Macro Av | Co | RL | AP | OE |
| LdSM [16] | | | | √ | | | | | | | | | | | | | |
| 3RC [35] | √ | | √ | | √ | √ | | | | | | | | | | √ | |
| MLNB-LD [29] | | | √ | | | | | | | | √ | √ | | | | | |
| BCC [38] | √ | | √ | | | | | | | | | | √ | | | | |
| RBRL [40] | √ | √ | | | | √ | | | | | | | | √ | √ | √ | |
| AEDC-MLSVM [20] | √ | | | | | | | | | | | | | √ | √ | √ | |
| SSSVM [41] | √ | | | | | | | | | | | | | | | | |
| AEML-LLSVM [19] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| BP-AEPML [21] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| NNMLInf [22] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| LCL-Net [42] | | | | √ | √ | √ | | | | | | | | | | | |
| SDLM [43] | | | | | | √ | | | | | | | | | | | |
| DCNet [44] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| ELIFT [32] | | √ | √ | | | | | | | | | | | | | | |
| ACkEL [48] | √ | √ | | | | | | | | | √ | | | √ | √ | √ | √ |
| AESAKNNS [33] | | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | | | | | |
| RkNN [29] | | | | √ | √ | | | | | | | √ | | | | | √ |
| MLSAkNN [31] | | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | | | | | |
| ML-BTC [36] | √ | √ | √ | | | | | | | | √ | √ | | | √ | | √ |
| BNCC [39] | √ | | | | | √ | | | | | √ | √ | | | | | |
| ML-KELM [26] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| ML-ELM-RBF [47] | √ | | | | | | | | | | | | | √ | √ | √ | √ |
| ML-AP-RBF-Lap-ELM [25] | √ | | | | √ | | | | | | | | | √ | | √ | √ |
| ML-CK-ELM [27] | √ | | | | | | | | | | | | | √ | √ | √ | √ |

Precision and recall can be used to calculate the weighted F1 metric. This metric is generally considered to be a better performance evaluation index than precision and recall.

$$\text{F1 - measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{17}$$

In order to better understand the role of evaluation metrics in multi-labeling, the evaluation metrics of the above SL algorithms are depicted in Table 8. where SA is subset precision, Acc is accuracy, Pre is precision, Re is recall, Micro Pre is micro-prediction, Macro Pre is macro-prediction, Micro Re is micro-recall, Macro Re is macro-recall, Macro Av is macro-average, Co is coverage, RL is Ranking Loss, AP is average precision, OE is One Error.

## 4.2 Public dataset

The main application areas of public datasets are media, biology, text, image, and chemistry, etc. Selected datasets can be downloaded from these three Web sites: http://mulan.sourceforge.net/datasets-mlc.html, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ and http://www.uco.es/kdis/mllresources/. In domain, there are many relevant datasets, mainly described as follows:

Multimedia: Birds cover audio data. Cal500 Contains information about music clips. Emotions covers data about music clips with emotional labels. Mediamill covers data about the concepts that appear in the video.

Text: Bibtex contains information about bibtex project metadata, enron contains data about the emails of Enron seniors, and medical, a dataset whose instances correspond to documents with a summary of a patient symptom history.

Image: Corel5k is a data set whose examples correspond to Corel images that have been segmented through standardized cutting. Scene contains information about the scene, which can be annotated in the following six contexts: mountain, city, beach, sunset, field, and fallen leaves. Flags contains information about national flags.

Biology: Two datasets are relevant to this domain. The first is Yeast, which contains information about gene function. A second data set corresponding to biology is genbase, which contains data on proteins.

**Table 9** Summary of datasets

| Datasets | Domain | *N* | *M* | *L* | *LD(D)* |
|---|---|---|---|---|---|
| Rcv1-v2 | Text | 804,414 | 500 | 103 | 0.031 |
| IMDB | Text | 120,919 | 1001 | 28 | 0.071 |
| Mediamill | Multimedia(Video) | 43,907 | 120 | 101 | 0.043 |
| Tmc2007 | Text | 28,596 | 49,060 | 22 | 0.098 |
| 20NG | Text | 19,300 | 1006 | 20 | 1 |
| Delicious | Text | 16,105 | 500 | 983 | 0.019 |
| Ohsumed | Text | 13,529 | 1002 | 23 | 0.072 |
| bibtex | Text | 7395 | 1836 | 159 | 0.015 |
| Reuters | Text | 6000 | 500 | 101 | 0.028 |
| Corel5k | Image | 5000 | 499 | 374 | 0.009 |
| Slashdot | Text | 3782 | 1079 | 22 | 0.053 |
| Yeast | Biology | 2417 | 103 | 14 | 0.303 |
| Scene | Image | 2407 | 294 | 6 | 0.179 |
| Image | Image | 2000 | 294 | 5 | 0.247 |
| Enron | Text | 1702 | 1001 | 53 | 0.064 |
| Medical | Text | 978 | 1449 | 45 | 0.028 |
| Genbase | Biology | 662 | 1186 | 27 | 0.046 |
| Birds | Multimedia(audio) | 645 | 258 | 19 | 0.053 |
| Emotions | Multimedia(music) | 593 | 72 | 6 | 0.311 |
| CAL500 | Multimedia(music) | 502 | 68 | 174 | 0.15 |
| Flags | Image | 194 | 10 | 7 | 0.485 |

The attributes of the data set can be counted from *N*: number of instances, *M*: number of features, *L*: number of class labels, *LD (D)*: label density.

In order to better understand the information of the datasets, some datasets are introduced in detail in Table 9, the table is listed below in descending order of the number of instances.

## 5 Next direction

The existing MLC literatures based on SL and SSL have been able to solve the classification problem very well, but there are still some serious problems for researchers to solve, for example, processing of complex concept drift, processing of complex correlations of label, processing of feature selection, and processing of class imbalance. The following will analyze these issues and serve as the future research directions of this article.

### 5.1 Processing of complex concept drift

Nowadays, data stream becomes more and more common and the imperative for online algorithms for mining transient and dynamic data is becoming more and more evident [135]. At present, there are few MLC algorithms to solve concept drift, which makes concept drift a worthy research direction. Since there are various types of concept drift, such as gradual drift, abrupt drift, repeated drift, etc., how to effectively detect concept drift has become an urgent challenge. Block-based and incremental update strategies are widely used in single-label algorithms and have achieved good results. The research group will decide to convert it into a binary classifier using BR method. The algorithm based on block and incremental update strategy is used to detect concept drift.

### 5.2 Processing of complex correlations of label

Existing classification methods simply consider the correlation between labels, but some labels have very complex relationships with each other. Some labels within a dataset have bidirectional relationships and multiple periodic dependencies. For example, the prediction of the "beach" category depends on the "city" value, while the prediction of the "city" category depends on the "beach" value [35]. This makes the correlation between labels important for the study of MLC problems under SL and SSL. Many algorithms only partially consider the complex label correlation problem, and effectively considering the label correlation can improve the classification performance.

### 5.3 Processing of feature selection

Feature selection is the process of data pre-processing. Algorithms can reduce complexity and improve prediction accuracy through feature selection. In general, the presence of redundant or irrelevant attributes may cause other problems such as poor classification performance and may have high computational and memory storage requirements [101]. Multi-label algorithms based on SL and SSL select a subset of features that contain highly relevant and non-redundant features can filter out redundant features to a large extent. Currently, the most basic feature selection methods can be broadly classified as packing, embedding, and filtering methods. However, in general, they must collect the complete set of features before feature selection starts. This has some limitations because in reality many features are dynamically changing. In view of this feature, online feature selection methods can be used to deal with the multi-label problem and solve the feature selection problem.

### 5.4 Processing of class imbalance

Most multi-label datasets have a serious class imbalance, which will seriously affect the classification performance [20]. The class imbalance data are divided two aspects: on the one hand, for a particular class label, the number of positive instances is significantly less than the number of negative instances. On the other hand, for a particular instance, the

number of relevant labels is usually less than the number of irrelevant labels [40]. Traditional classifiers are more suitable for the classification of balanced data, because the classification performance will decline sharply when classes are imbalanced in the multi-label data. Therefore, it is important to take SL and SSL as the next step to deal with the class imbalance problem.

## 6 Summary

This paper introduces the existing MLC algorithms based on SL learning and SSL. At the same time, it summarizes algorithms of practical application fields such as multi-label image and text classification, and summarizes the involved algorithms from multiple aspects through images and tables. Then the evaluation metrics and public datasets of multi-label are briefly introduced. Finally, we propose the next research directions based on the current challenges faced by MLC.

By reviewing supervised and semi-supervised learning algorithms for multi-label classification, we can understand that more and more algorithms consider the correlation between labels and it can improve the classification performance of the algorithms in supervised learning algorithms. Semi-supervised learning algorithms are significantly more important when there are labeled and unlabeled data in the dataset. The inductive methods are optimized for the classification model, while the transductive methods are optimized directly for the prediction. Multi-label classification algorithms can be applied in many real scenes, mainly images and text. The image field is mainly divided into medicine and remote sensing, and the text field is mainly divided into emotion and medical biology. If an algorithm considers only one evaluation metric alone it may not yield as good results as another metric, but this does not mean that the metric has no role at all in the evaluation and it can be chosen to be used in combination with more sensitive evaluation metrics. In general, multiple evaluation metrics should be provided when measuring algorithm performance, rather than allowing performance to be determined by a single evaluation metric. With the rapid development of big data, more and more data are generated in our daily life, multi-label classification algorithms are becoming more and more important, but they also face many challenges. We can continue to study the problems of complex concept drift, complex label relationships, feature selection and class imbalance.

**Acknowledgements** This work is supported by the National Nature Science Foundation of China (62062004), the Ningxia Natural Science Foundation Project (2022AAC03279).

**Funding** The National Nature Science Foundation of China (Grant no. 62062004); The Ningxia Natural Science Foundation Project (Grant no. 2022AAC03279).

## Declarations

**Conflict of interest** All authors have no financial or non-financial interests directly or indirectly related.

## References

1. Zhang X, Han M, Wu H et al (2021) An overview of complex data stream ensemble classification. J Intell Fuzzy Syst 1–29
2. Ma J, Zhang H, Chow TWS (2021) Multilabel classification with label-specific features and classifiers: a coarse- and fine-tuned framework. IEEE Trans Cybern 51(2):1028–1042
3. Read J, Pfahringer B, Holmes G et al (2011) Classifier chains for multi-label classification. Mach Learn 85(3):333–359
4. Tsoumakas G, Vlahavas I (2007) Random k-Labelsets: an ensemble method for multilabel classification. In: Proc of the 18th European conference on machine learning, Springer Berlin Heidelberg. Lecture Notes in Computer Science, Warsaw, pp 406–417
5. Oliveira E, Ciarelli PM, Badue C et al (2008) A comparison between a KNN based approach and a PNN algorithm for a multi-label classification problem. In: Proc of the eighth international conference on intelligent systems design applications. IEEE, Kaohsiung, pp 628–633
6. Clare A, King RD (2001) Knowledge discovery in multi-label phenotype data. In: Proc of the European conference on principles of data mining and knowledge discovery. Lecture notes in computer science, Freiburg, pp 42–53
7. Li X, Wang L, Sung E (2004) Multi-label SVM active learning for image classification. In: Proc of the image processing. IEEE, Singapore, pp 2207–2210
8. Sapozhnikova EP (2009) Multi-label classification with ART neural networks. In: Proc of the 2009 second international workshop on knowledge discovery and data mining. IEEE, Moscow, pp 144–147
9. Tsoumakas G, Katakis I (2009) Multi-label classification: an overview. Int J Data Warehouse Min 3(3):1–13
10. Moyano JM, Gibaja EL, Cios KJ et al (2018) Review of ensembles of multi-label classifiers: models, experimental study and prospects. Inf Fusion 44:33–45
11. Zheng X, Li P, Chu Z et al (2020) A survey on multi-label data stream classification. IEEE Access 8:1249–1275
12. Engelen JE, Hoos HH (2020) A survey of semi-supervised learning. Mach Learn 109(2):373–440
13. Li P, Wang H, Bhm C et al (2020) Online semi-supervised multi-label classification with label compression and local smooth regression. In: Proc of the twenty-ninth international joint conference on artificial intelligence, Yokohama, pp 1359–1365
14. Wang Z, Wang T, Wan B et al (2020) Partial classifier chains with feature selection by exploiting label correlation in multi-label classification. Entropy 22(10):1–22
15. Bezembinder EM, Wismans LJJ, Berkum ECV (2017) Constructing multi-labelled decision trees for junction design using the predicted probabilities. In: Proc of the 20th IEEE international conference on intelligent transportation systems. IEEE, Yokohama, pp 1–7
16. Majzoubi M, Choromanska A (2019) LdSM: logarithm-depth streaming multi-label decision trees. In: Proc of the 23rd

international conference on artificial intelligence and statistics, Palermo, pp 4247–4257
17. Moral-García S, Mantas CJ, Castellano JG et al (2020) Non-parametric predictive inference for solving multi-label classification. Appl Soft Comput 88
18. Yang Y, Ding M (2019) Decision function with probability feature weighting based on Bayesian network for multi-label classification. Neural Comput Appl 31(9):4819–4828
19. Sun Z, Hu K, Hu T et al (2018) Fast multi-label low-rank linearized SVM classification algorithm based on approximate extreme points. IEEE Access 42319–42326
20. Sun Z, Liu X, Hu K et al (2020) An efficient multi-label SVM classification algorithm by combining approximate extreme points method and divide-and-conquer strategy. IEEE Access 8:170967–170975
21. Wang X, Guo Z, Wang X et al (2019) A fast neural network multi-label classification algorithm based on approximate extreme points. In: 2019 5th international conference on big data computing and communications (BIGCOM)
22. Wang X, Guo Z, Wang X, et al (2019) NNMLInf: social influence prediction with neural network multi-label classification. In: Proc of the ACM turing celebration conference, vol 106. ACM, Chengdu, pp 1–5
23. Bello M, Gonzalo N, Ricardo S, et al (2020) Deep neural network to extract high-level features and labels in multi-label classification problems. Neurocomputing 413
24. Sumbul G, Begüm D (2020) A deep multi-attention driven approach for multi-label remote sensing image classification. IEEE Access 8:95934–95946
25. Xu X, Shan D, Li S et al (2019) Multi-label learning method based on ML-RBF and laplacian ELM. Neurocomputing
26. Luo F, Guo W, Yu Y et al (2017) A multi-label classification algorithm based on Kernel extreme learning machine. Neurocomputing 260(Oct 18):313–320
27. Rezaei M, Eftekhari M, Movahed FS (2020) ML-CK-ELM: an efficient multi-layer extreme learning machine using combined kernels for multi-label classification. Scientia Iranica (6)
28. Rr A, Mea B, Sm C Regularizing extreme learning machine by dual locally linear embedding manifold learning for training multi-label neural network classifiers. Eng Appl Artif Intell 97
29. Sadhukhan P, Palit S (2020) Multi-label learning on principles of reverse k-nearest neighbourhood. Expert Syst
30. Roseberry M, Krawczyk B, Cano A (2019) Multi-label punitive kNN with self-adjusting memory for drifting data streams. ACM Transactions on Knowledge Discovery from Data
31. Roseberry M, Krawczyk B, Djenouri Y et al (2021) Self-adjusting k nearest neighbors for continual learning from multi-label drifting data streams. Neurocomputing
32. Wei X, Yu Z, Zhang C et al (2018) Ensemble of label specific features for multi-label classification. In: Proc of the 2018 IEEE international conference on multimedia and expo. IEEE, San Diego, pp 1–6
33. Alberghini G, Junior SB, Cano A (2022) Adaptive ensemble of self-adjusting nearest neighbor subspaces for multi-label drifting data streams. Neurocomputing 481:228–248
34. Lee CH (2018) Multi-label classification of documents using fine-grained weights and modified co-training. Intell Data Analysis 22(1):103–115
35. Lotf H, Ramdani M (2020) Multi-label classification: a novel approach using decision trees for learning label-relations and preventing cyclical dependencies: Relations Recognition and Removing Cycles (3RC). In: Proc of the 13th international conference on intelligent systems: theories and applications. ACM, Rabat, pp 1–6
36. Law A, Ghosh A (2021) Multi-label classification using binary tree of classifiers. IEEE Trans Emerg Topics Comput Intell 6(3):677–689
37. Kim H, Park J, Kim D, Lee J (2020) Multilabel naïve Bayes classification considering label dependence. Pattern Recogn Lett 136:279–285
38. Jiménez BR, Morales EF, Escalante HJ (2018) Bayesian chain classifier with feature selection for multi-label classification 232–243
39. Wang R, Ye S, Li K et al (2020) Bayesian network based label correlation analysis for multi-label classifier chain. Inf Sci 554(8)
40. Wu G, Zheng R, Tian Y, Liu D (2020) Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification. Neural Netw 122:24–39
41. Koda S, Zeggada A, Melgani F et al (2018) Spatial and structured SVM for multilabel image classification. IEEE Trans Geosci Remote Sens 56(10):5948–5960
42. Chen B, Li J, Lu G et al (2019) Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification. IEEE J Biomed Health Inform 24(8):2292–2302
43. Li J, Fu G, Chen Y et al (2020) A multi-label classification model for full slice brain computerised tomography image. BMC Bioinform. https://doi.org/10.1186/s12859-020-3503-0
44. He J, Li C, Ye J, Qiao Y, Gu L (2021) Multi-label ocular disease classification with a dense correlation deep neural network. Biomed Signal Process Control 63:102167
45. Krishna GS, Prakash N (2021) Deep learning for efficient and multi-labelled classification of synthetic aperture radar images. Evolv Syst, 1–14
46. Zhou C, Chen H, Jing Z, et al (2021) Multi-label graph node classification with label attentive neighborhood convolution. Expert Syst Appl
47. Nan Z, Ding S, Jian Z (2016) Multi layer ELM-RBF for multi-label learning. Appl Soft Comput 43:535–545
48. Wang R, Kwong S, Jia Y et al (2021) Active k-labelsets ensemble for multi-label classification. Pattern Recogn 109
49. Wang R, Kwong S, Jia Y et al (2018) Mutual information based K-labelsets ensemble for multi-label classification. In: Proceedings of the 2018 IEEE international conference on fuzzy systems. IEEE, Rio de Janeiro, pp 1–7
50. Li GZ, Yang JY, Lu WC et al (2008) Improving prediction accuracy of drug activities by utilising unlabelled instances with feature selection. Int J Comput Biol Drug Des 1(1):1–13
51. Zhang M-L, Zhou Z-H (2007) Ml-knn: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
52. Li GZ, You M, Ge L et al (2010) Feature selection for semi-supervised multi-label learning with application to gene function analysis. In: Proceedings of the first ACM international conference on bioinformatics and computational biology, pp 354–357
53. Xu M, Sun F, Jiang X (2014) Multi-label learning with co-training based on semi-supervised regression. In: Proceedings of the IEEE international conference on security, pp 175–180
54. Gharroudi O, Elghazel H, Aussem A (2017) A semi-supervised ensemble approach for multi-label learning. In: Proceedings of the IEEE international conference on data mining workshops. IEEE, pp 1197–1204
55. Ai Q, Kang Y, Wang A et al (2020) An effective semi-supervised multi-label least squares twin support vector machine. IEEE Access 8:213460–213472
56. Gharebaghi F, Amiri A (2021) LP-MLTSVM: laplacian multi-label twin support vector machine for semi-supervised classification. IEEE Access 10:13738–13752
57. Zhan W, Zhang ML (2017) Inductive semi-supervised multi-label learning with co-training. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1305–1314

58. Settouti N, Douibi K, Bechar MEA et al (2019) Semi-supervised learning with collaborative bagged multi-label K-nearest-neighbors. Open Comput Sci 9(1):226–242
59. Chu Z, Li P, Hu X (2019) Co-training based on semi-supervised ensemble classification approach for multi-label data stream. In: Proceedings of the 2019 IEEE international conference on big knowledge (ICBK). IEEE, pp 58–65
60. Alalga A, Benabdeslem K, Taleb N Soft-constrained Laplacian score for semi-supervised multi-label feature selection. Knowledge Inf Syst
61. Santos AM, Canuto A (2014) Applying the self-training semi-supervised learning in hierarchical multi-label methods. In: Proceedings of the international joint conference on neural networks, pp 872–879
62. Santos AM, Canuto AMP (2012) Using semi-supervised learning in multi-label classification problems. In: Proceedings of the 2012 international joint conference on neural networks, pp 1–8
63. Rodrigues FM, Canuto A, Santos AM (2014) Confidence factor and feature selection for semi-supervised multi-label classification methods. In: Proceedings of the international joint conference on neural networks, pp 864–871
64. Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336
65. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
66. Zhao C, Zhai S (2016) Minimum variance semi-supervised boosting for multi-label classification. In: Proceedings of the IEEE global conference on signal & information processing. IEEE
67. Ahmed MS, Khan L, Oza NC et al (2010) Multi-label asrs dataset classification using semi-supervised subspace clustering. In: Proceedings of the 2010 conference on intelligent data understanding, pp 285-299
68. Mehrkanoon S, Suykens J (2016) Multi-label semi-supervised learning using regularized kernel spectral clustering. In: Proceedings of the international joint conference on neural networks. IEEE, pp 4009–4016
69. Amorim WP, Falcao AX, Papa JP (2018) Multi-label semi-supervised classification through optimum-path forest. Inf Sci 465:86–104
70. Pham TN, Nguyen VQ, Tran VH et al (2017) A semi-supervised multi-label classification framework with feature reduction and enrichment. J Inf Telecommun 1(2):141–154
71. Ha QT, Pham TN, Nguyen VQ et al (2018) A new text semi-supervised multi-label learning model based on using the label-feature relations. In: Proceedings of the computational collective intelligence—10th international conference, pp 403–413
72. Huang J, Xu L, Qian K et al (2021) Multi-label learning with missing and completely unobserved labels. Data Min Knowl Disc 35(3):1061–1086
73. Jing L, Yang L, Yu J et al (2015) Semi-supervised low-rank mapping learning for multi-label classification. In: Proceedings of the 2015 IEEE conference on computer vision and pattern recognition. IEEE, pp 1483–1491
74. Sheng L, Yun F (2017) Robust multi-label semi-supervised classification. In: Proceedings of the 2017 IEEE international conference on big data (big data). IEEE, pp 27–36
75. Sun L, Feng S, Lyu G, Lang C (2019) Robust semi-supervised multi-label learning by triple low-rank regularization. In: Proceedings of the advances in knowledge discovery and data mining, pp 269–280
76. Sun L, Lyu G, Feng S et al (2021) Beyond missing: weakly-supervised multi-label learning with incomplete and noisy labels. Appl Intell 51(3):1552–1564
77. Zha ZJ, Tao M, Wang J et al (2009) Graph-based semi-supervised learning with multi-label. J Vis Commun Image Represent 20(2):97–103
78. Li T, Yan S, Mei T, Kweon I (2009) Local-driven semi-supervised learning with multi-label. In: Proceedings of the 2009 IEEE international conference on multimedia and expo, pp 1508–1511
79. Bao BK, Li T, Yan S (2009) Hidden-concept driven image decomposition towards semi-supervised multi-label image annotation. In: Proceedings of the first international conference on internet multimedia computing and service, pp 17–24
80. Jiang JQ (2012) Predicting protein function by multi-label correlated semi-supervised learning. IEEE/ACM Trans Comput Biol Bioinf 9(4):1059–1069
81. Liu Y, Nie F, Gao Q (2018) Nuclear-norm based semi-supervised multiple labels learning. Neurocomputing 275:940–947
82. Ghosh A, SeKhar CC (2017) Label correlation propagation for semi-supervised multi-label learning. In: Proceedings of the pattern recognition and machine intelligence, pp 52–60
83. Tan Q, Yu Y, Yu G et al (2017) Semi-supervised multi-label classification using incomplete label information. Neurocomputing 260(18):192–202
84. Behpour S, Xing W, Ziebart BD (2018) ARC: adversarial robust cuts for semi-supervised and multi-label classification. In: Proceedings of the IEEE/CVF conference on computer vision & pattern recognition workshops. IEEE, pp 2704–2711
85. Gao K, Zhang J, Zhou C (2019) Semi-supervised graph embedding for multi-label graph node classification. 555–567
86. Tang C, Liu X, Wang P et al (2019) Adaptive hypergraph embedded semi-supervised multi-label image annotation. IEEE Trans Multimedia 21(11):2837–2849
87. Wu F, Wang Z, Zhang Z et al (2015) Weakly semi-supervised deep learning for multi-label image annotation. IEEE Trans Big Data 1(3):1–1
88. Wang X, Feng S, Lang C (2018) Semi-supervised dual low-rank feature mapping for multi-label image annotation. Multimedia Tools Appl
89. Zhang Q, Zhong G, Dong J (2021) A graph-based semi-supervised multi-label learning method based on label correlation consistency. Cognitive Comput 1–10
90. Song Z, Meng Z, Zhang Y, King I (2021) Semi-supervised multi-label learning for graph-structured data. In: The 30th ACM international conference on information and knowledge management, pp 1723–1733
91. Boulbazine S, Cabanes G, Matei B et al (2018) Online semi-supervised growing neural gas for multi-label data classification. In: Proceedings of the 2018 international joint conference on neural networks, pp 1–8
92. Li D, Dick S (2018) Residential household non-intrusive load monitoring via graph-based multi-label semi-supervised learning. In: IEEE transactions on smart grid, pp 1–1
93. Hu S, Miao D, Pedrycz W (2022) Multi granularity based label propagation with active learning for semi-supervised classification. Expert Syst Appl 192:116276
94. Bo W, Tu Z, Tsotsos JK (2014) Dynamic label propagation for semi-supervised multi-class multi-label classification. In: Proceedings of the IEEE international conference on computer vision, vol 68, pp 14–23
95. Dharmadhikari SC, Ingle M, Kulkarni P (2012) Semi supervised learning based text classification model for multi label paradigm. In: Proceedings of the signal processing and information technology—second international joint conference, pp 178–184

96. Lucena D, Prudencio R (2015) Semi-supervised multi-label k-nearest neighbors classification algorithms. In: Proceedings of the 2015 Brazilian conference on intelligent systems. IEEE, pp 49–54
97. Gang C, Song Y, Fei W et al (2008) Semi-supervised multi-label learning by solving a sylvester equation. In: Proceedings of the Siam international conference on data mining, pp 410–419
98. Cevikalp H, Benligiray B, Gerek ON (2019) Semi-supervised robust deep neural networks for multi-label image classification. Pattern Recogn 100:107164
99. Guan Q, Huang Y (2018) Multi-label chest X-ray image classification via category-wise residual attention learning. Pattern Recogn Lett 130:259–266
100. Chougrad H, Zouaki H, Alheyane O (2020) Multi-label transfer learning for the early diagnosis of breast cancer. Neurocomputing 392:168–180
101. Liu J, Lin Y, Li Y et al (2018) Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recogn 84:273–287
102. Cai J, Sun W, Guan J et al (2020) Multi-ECGNet for ECG arrythmia multi-label classification. IEEE Access 8:110848–110858
103. Ou X, Gao L, Quan X et al (2022) BFENet: a two-stream interaction CNN method for multi-label ophthalmic diseases classification with bilateral fundus images. Comput Methods Programs Biomed 219:106739
104. Xu S, Yang X, Guo J et al (2020) CXNet-m3: a deep quintuplet network for multi-lesion classification in Chest X-ray Images via multi-label supervision. IEEE Access 8:98693–98704
105. Hua Y, Mou L, Zhu XX (2020) Relation network for multilabel aerial image classification. IEEE Trans Geosci Remote Sens 58:4558–4572
106. Alshehri A, Bazi Y, Ammour N et al (2019) Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery. IEEE Access 7:119873–119880
107. Chaudhuri B, Demir B, Chaudhuri S et al (2018) Multi-label remote sensing image retrieval using a semisupervised graph-theoretic method. IEEE Trans Geosci Remote Sens 56(2):1144–1158
108. Dai OE, Demir B, Sankur B, Bruzzone L (2018) A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sensing 11(7):2473–2490
109. He H, Xia R (2018) Joint binary neural network for multi-label learning with applications to emotion classification. 11108:250–259
110. Yu J, Luís M, Jiang J et al (2018) Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In: Proceedings of the 2018 conference on empirical methods in natural language processing
111. Fei H, Ji D, Zhang Y et al (2020) Topic-enhanced capsule network for multi-label emotion classification. IEEE/ACM Trans Audio Speech Lang Process 28:1839–1848
112. Alzu'Bi S, Badarneh O, Hawashin B et al (2019) Multi-label emotion classification for Arabic tweets. In: Proc of the 2019 sixth international conference on social networks analysis. IEEE, Granada, pp 499–504
113. Bravo-Marquez F, Frank E, Mohammad SM et al (2016) Determining word-emotion associations from tweets by multi-label classification. In: Proc of the IEEE/WIC/ACM international conference on web intelligence ACM, . Omaha, NE, pp 536–539
114. Kim Y, Lee H, Jung K (2018) AttnConvnet at SemEval-2018 Task 1: attention-based convolutional neural networks for multi-label emotion classification. In: Proc of the 12th international workshop on semantic evaluation. Association for Computational Linguistics, New Orleans, pp 141–145
115. Mulki H, Ali CB, Haddad H et al (2018) Tw-StAR at SemEval-2018 Task 1: preprocessing impact on multi-label emotion classification. In: Proc of the SemEval-2018. Association for Computational Linguistics, New Orleans, pp 167–171
116. Alhuzali H, Ananiadou S (2021) SpanEmo: casting multi-label emotion classification as span-prediction. Comput Sci
117. Hyun J, Bae B, Cheong Y (2020) [CL-AFF Shared Task] multi-label text classification using an emotion embedding model. In: Proc of the 3rd workshop of affective content analysis. CEUR Workshop Proceedings, New York, pp 169–178
118. Ying W, Xiang R, Lu Q (2019) Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In: Proc of the 5th workshop on noisy user-generated text. Association for Computational Linguistics, Hong Kong, pp 316–321
119. Ameer I, Sidorov G, Gomez-Adorno H et al (2022) Multi-label emotion classification on code-mixed text: data and methods. IEEE Access 10:8779–8789
120. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N (2018) Multi-label classification of patient notes: case study on ICD code assignment. AAAI Workshops, pp 409–416
121. Du J, Chen Q, Peng Y, Xiang Y et al (2019) ML-Net: multi-label classification of biomedical texts with deep neural networks. J Am Med Inform Assoc 26(11):1279–1285
122. Glinka K, Wozniak R, Zakrzewska D (2017) Improving multi-label medical text classification by feature selection. In: Proc of IEEE international conference on enabling technologies: infrastructure for collaborative enterprises. IEEE, Poznan, pp 176–181
123. Hughes M, Li I, Kotoulas S, Suzumura (2017) Medical text classification using convolutional neural networks. Comput Sci
124. Yogarajan V, Montiel J, Smith T, Pfahringer B (2020) Seeing the whole patient: using multi-label medical text classification techniques to enhance predictions of medical codes. Comput Sci
125. Wasim M, Mahmood W, Asim MN et al (2019) Multi-label question classification for factoid and list type questions in biomedical question answering. IEEE Access 7:3882–3896
126. D'Anniballe VM, Tushar FI, Faryna K et al (2022) Multi-label annotation of text reports from computed tomography of the chest, abdomen, and pelvis using deep learning. BMC Med Inform Decis Mak 22(1):1–12
127. Oramas S, Nieto O, Barbieri F et al (2017) Multi-label music genre classification from audio, text, and images using deep features 23–30
128. Zhao G, Xu J, Zeng Q et al (2019) Review-driven multi-label music style classification by exploiting style correlations. In: Proc of the 2019 conference of the north american chapter of the association for computational linguistics. Association for Computational Linguistics, Minneapolis, pp 2884–2891
129. Ma Q, Yuan C, Zhou W, Han J, Hu S (2020) Beyond statistical relations: integrating knowledge relations into style correlations for multi-label music style classification. In: Proc of the 13th international conference on web search and data mining. ACM, Houston, pp 411–419
130. Kim E, On K, Kim J et al (2018) Temporal attention mechanism with conditional inference for large-scale multi-label video classification. In: Proc of the European conference on computer vision. Lecture Notes in Computer Science, Munich, pp 306–316
131. Karagoz GN, Yazici A, Dokeroglu T, Cosa A (2020) Analysis of multiobjective algorithms for the classification of multi-label video datasets. IEEE Access 8:163937–163952
132. Araujo A, Negrevergne B, Chevaleyre Y et al (2018) Training compact deep learning models for video classification using

circulant matrices. In: Proc of European conference on computer vision. Lecture Notes in Computer Science, Munich, pp 271–286
133. Jiang B, Zhou L, Lin L et al (2019) A real-time multi-label classification system for short videos. In: Proc of 2019 IEEE international conference on image processing. IEEE, Taipei, pp 534–538
134. Wu H, Li M, Liu Y et al (2022) Transtl: spatial-temporal localization transformer for multi-label video classification. In: ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1965–1969
135. Büyükçakır A, Bonab H, Can F (2018) A novel online stacked ensemble for multi-label stream classification. In: Proc of the 27th ACM international conference on information and knowledge management, pp 1063–1072