Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **6**, Number **2**, May **2021**

Multi-Label Classification Review and Opportunities

Wei Weng

College of Computer and Information Engineering
Xiamen University of Technology
Xiamen, 361024, China
xmutwei@163.com

Yu-Wen Li

School of Instrument Science and Engineering
Southeast University
Nanjing, 210096, China
liyuwen@seu.edu.cn

Jing-Hua Liu

College of Computer Science and Technology
Huaqiao University
Xiamen, 361021, China
xmujhliu@163.com

Shun-Xiang Wu

Department of Automation
Xiamen University
Xiamen, 361005, China
sxwu@xmu.edu.cn

Chin-Ling Chen

School of Computer and Information Engineering
Xiamen University of Technology
Xiamen 361005, China
School of Information Engineering
Changchun Sci-Tech University
Changchun 130600, China
Department of Computer Science and Information Engineering
Chaoyang University of Technology
Taichung 41349, Taiwan
clc@mail.cyut.edu.tw

Corresponding Author: Chin-Ling Chen

clc@mail.cyut.edu.tw

Received Januray 2021; revised March 2021

Abstract. *Multi-label classification originated from text classification and has became one of the most widely studied machine learning frameworks. After nearly twenty years of development, many multi-label classification models have been produced. In this paper, the representative algorithms are introduced and reviewed. On the other hand, in recent years, the volume of multi-label data has become larger, and the features and labels have become higher dimensions, so the multi-label algorithms have also generated new trends. The main research directions focus on label specific feature, label correlation, and dimension compression. This paper discusses these studies in detail. Different from other review works, this paper discusses the evolution and improvement process of various algorithms and summarizes the main research directions in recent years. The content summarized in this paper can provide a more comprehensive perspective for related researchers to understand the main research contents in this direction, and inspire new research means and methods.*

Keywords: Multi-label classification, label-specific feature, label correlation, dimension compression

1. **Introduction.** The traditional single-label classification method allocates an exclusive class label for each instance from a limited class label set, which assures mutual exclusion among different class labels and prevents the coexistence of different class labels in the same instance. Single-label classification is divided into binary classification and multi-class classification [51]. If the class label set only has two elements, it is called binary classification; otherwise, it is called multi-class classification. A single-label problem is a simplified model of the real world. Objects in real life often have complicated semantics and they require multi-labels for description. For example, genes often have multiple functions, such as metabolism, transcription, and protein synthesis [4, 14]; one article can involve multiple themes [52, 65, 111], and the contents of one picture may cover several sceneries [15, 63, 72]. Multi-label data exist extensively in the real world, and the total number of labels varies significantly among different types of application problems, ranging from dozens to thousands. For instance, the number of labels in photo tagging application can reach as high as tens of thousands [91]; and the number of labels for text classification task can reach millions [13].

Multi-label data inherit rich semantics, which is helpful to process many practical problems effectively. For example, annotating multi-label for the documents on Internet can increase the accuracy of personalized recommendation [1]; Multi-labeling for photos uploaded onto the network can increase text-based image retrieval accuracy [64]; in social network analysis, giving multi-labels onto nodes is conducive to overlapping community discovery, which helps to disclose dynamic features of real networks [78]. To disclose semantic information of such data and meet the demands of applications, it is urgent to investigate multi-label classification methods based on machine learning effectively [21]. Of course, the simplest way to solve multi-label problems is to divide a multi-label problem into several traditional single-label problems and each label is processed independently. However, it neglects relations among labels and is inconvenient to consider relations between features and labels in the learning process. A dedicated machine learning model, namely multi-label classification, was designed to cope with such tasks [112].

In multi-label classification, the predicted labels might be any subset of the full label set for any prediction instance, which is an exponential complicated problem. For example, the number of possible label sets might be millions if there are a total of 20 labels. Hence, multi-label classification is an extremely challenging task, especially for problems with tens of thousands of labels.

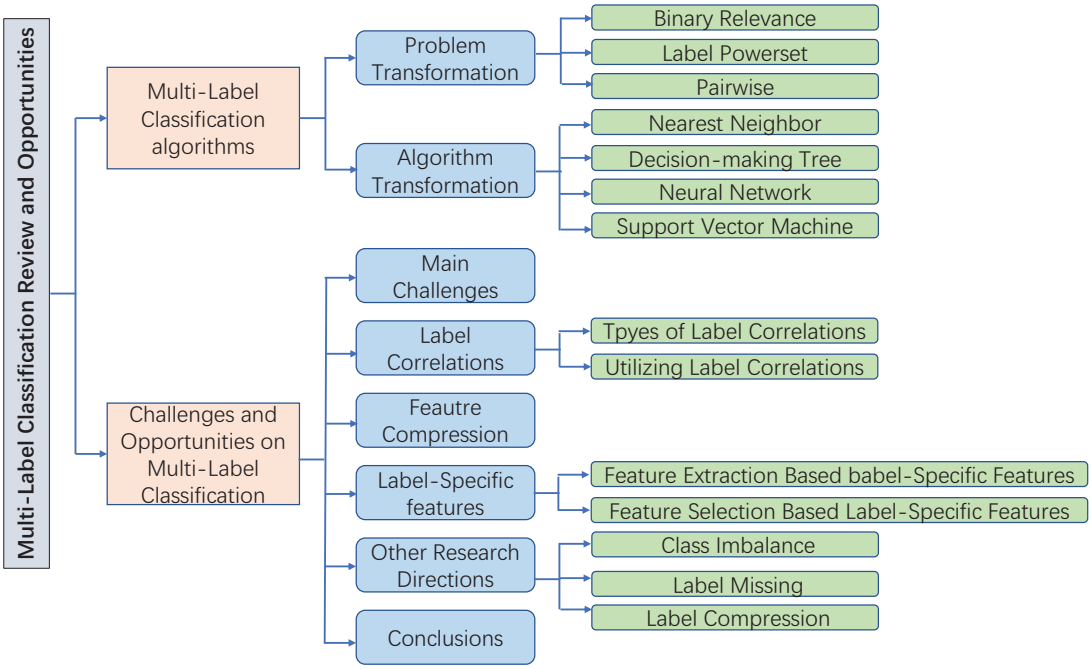

FIGURE 1. The structure of this paper

We will briefly review the main categories, improvement strategies, the latest challenges and opportunities about multi-label classification. It is worth mentioning that the metrics for evaluating the performance of multi-label models can be found in ref. [112]. The rest of this article is organized as follows. Section 2 reviews the main algorithm categories for multi-label classification, and the characteristics of various algorithms in each category and their evolution approaches. Section 3 presents the latest challenges, research progress and opportunities. Section 4 provides our conclusions. The structure of this paper is shown in Figure 1.

2. **Multi-Label Classification Algorithm.** Multi-label classification is originated from text classification [65, 79]. Research progresses of multi-label classification are driven by the need for abundant high-efficiency document retrieval on the Internet. With the continuous development of information acquisition and data transmission, massive multi-label data were accumulated continuously. Therefore, Researches on multi-label classification have been expanded to multiple domains gradually in recent years, including image and video retrieval, prediction of gene functions, music emotion classification, etc. Researchers have proposed various multi-label classification models, all of which process multi-label classification problems through problem transformation or algorithm transformation [112]. Problem transformation is to transform the multi-label problem into one or several single-label subproblems and then these subproblems are solved directly by mature single-label algorithms. Therefore, problem transformation is independent of a specific algorithm and an appropriate algorithm can be chosen according to practical situations. Algorithm transformation adopts another method: the existing single-label algorithm is expanded to be applied to multi-label data directly. For the convenience of description, relevant notations have to be defined first. Let $\mathcal{L} = \{y_1, y_2, \cdots, y_q\}$ be the full set of q labels, and the multi-label data covering n samples is expressed as $\mathcal{D} = \{(x_i, Y_i)|1 \leq i \leq n\}$, where $Y_i \in \mathcal{L}$ is the set of all annotated labels for the instance $\mathbf{x}_i$. Y_i is often expressed by the logic vector $\mathbf{c}_i = [c_{i1}, c_{i2}, \cdots, c_{iq}]$. If y_j is the label belonging to $\mathbf{x}_i$, there is $c_{ij} = 1$; otherwise, $c_{ij} = 0$.

2.1. **Problem Transformation Algorithm.** At present, the problem transformation algorithm is mainly based on the following three strategies: 1) binary relevance (BR) transformation method, 2) label powerset (LP) transformation method, and 3) pairwise methods (PW).

1) **BR transformation method**

BR is the most representative problem transformation model and it establishes the binary classification model for each label. The basic BR model [12] ignores label correlations. To exploit such correlations between labels in the BR framework, the common way is to add labels into the original features as extra features and then construct corresponding classifiers. These improved BR models can be divided into two types. The first type is to construct a two-layer BR. The first layer shows the same practices as the original BR. That is, it constructs binary classifiers for each label based on the original features. In the second layer, outputs of the first layer are added to the original features as extra features. Then, based on those augmented features, binary classifiers are learned for each label. The outputs of the second layer are used as the predicting labels. The BR model which uses this paradigm is called the stacking based BR model. The second type is to connect all binary classifiers into a chain, and a classifier on the chain adds outputs of all previous classifiers into the original features as extra features. Such a framework is called a classifier chain model. In the following, the BR model, stacking based BR model, and classifier chain models are introduced in detail.

- Basic BR model

In 2004, the basic BR model was firstly proposed by Boutell et al. [12]. It transforms a multi-label problem into several binary classification subproblems and each label is corresponding to one binary classification problem. For each instance in the training set, if the label is absent, this instance is considered as a negative instance; otherwise, it is a positive one. Obviously, all subproblems share the same feature space, but the label space is different. BR has two obvious advantages: one is that the computational complexity is linear with the number of labels; the other is that each label can be processed in parallel. Because of these advantages, many algorithms adopt the strategy of building a classifier for each label. Although the BR algorithm is simple, intuitive, and widely used, it often suffers from poor predictive performance because it ignores the relationship between labels.

- Stacking based BR model

To exploit label correlation, Godbole et al. [28] introduced the stacking-based learning strategy [92] in ensemble learning into the BR algorithm. During training, the stacking-based BR model established two layers of BR models: the first layer is the base level, which is the same as the traditional BR learning process and allocates the corresponding binary classification model for each label. In the second layer of meta-level, predicting labels of all binary classification models in the base level are added into the original feature spaces and then each label is learned again on those expanded features, getting the corresponding binary classification model. Obviously, such stacking based BR algorithm assumes that any label is related to all labels, which, however, is invalid in most cases. In other words, a label often introduces noises by using all outputs in the first layer as its new features. Therefore, some studies are engaged in selecting partial outputs of the first layer as new features to train the binary classifiers in the second layer, such as algorithm BR+ [22] and stacking based algorithm BR based on pruning [84].

- Classifier chain

J. Read et al [74]. proposed the classifier chains (CC) model for the first time. CC puts all labels in a random sequence. In this sequence, each label learns the corresponding binary classifier. Different from the BR algorithm, labels in the chain add binary classifier output of all previous labels into the original feature space as new features for training. Since the learning and test process of these binary classifiers has a certain sequence, it is called the classifier chain. CC model has two evident disadvantages. One is that the effect of classifiers is influenced greatly by the sequence of labels and different sequences bring significantly different classification effects. The other is that the current label may not be related to labels in the previous part of the sequence, so it might introduce noises by using outputs of all previous labels, thus decreasing performances of classifiers. In the same literature, J. Read proposed the ensemble framework (ECC) for CC. in ECC, the mean prediction results of the CC model of multiple random label sequences were taken, which could solve influences of random label sequence on the classification to some extent. However, this multiplied computing expenses. Dembczyński et al [24]. proposed the probabilistic classifier chains (PCC) model and pointed out that classifier chain based on loss functions like Hamming or Rank could be gained from the conditional joint distribution of labels. In addition to improving the effect of the CC algorithm based on probability deduction, there are some other methods engaged to find the optimal or the better label sequences. For example, GA-PartCC [29] searched the optimal label sequences by using genetic algorithms and OOCC [83] finds the optimal label sequence for each sample from the k-nearest neighbor.

2) **LP transformation method**

LP transformation method [85, 89] is to view the combination (subset) of all labels in a training set as a class and then transform a multi-label problem into a single-label multi-class problem. After a classifier is gained from learning, an unseen instance is an input and a class is an output. This class is corresponding to one label set which covers all labels the instance belongs to. LP transformation method also has some evident disadvantages. Firstly, this method can only predict label sets that occur in the training set. Moreover, there might be many label sets when there are a lot of labels. Therefore, many sets might possess a few corresponding instances, resulting in class imbalance. These problems not only increase the time cost of learning but also reduce the model effect. To solve these problems, RAkEL [85] and EPS [75] are two famous modified algorithms.

The Random k label sets (RAkEL) algorithm divides m subsets containing k labels from all labels randomly and then trains each label set by the LP transformation method. By using appropriate m and k, RAkEL can get significantly more accurate performance than BR and LP. Furthermore, RAkEL also decreases the number of models that have to be learned. During prediction, all learners gained by the LP method are used to predict the unseen instance and all prediction results are voted or averaged. Hence, RAkEL is a kind of ensemble learning method.

The EPS algorithm is also proposed to solve the shortcomings of LP. Firstly, EPS counts the number of the instance ($Count(R_i)$) related with the labelset (R_i). If $Count(R_i)$ is higher than the appointed threshold, all instances related to R_i are added into the training set. Secondly, if $Count(R_i)$ is lower than the threshold, the instances related to the subsets of R_i is continue to be counted. If the frequency of a subset is higher than the appointed threshold, samples corresponding to this subset are added to the training set corresponding to R_i. In this way, LP method is trained on the corresponding training set of R_i. Meanwhile, EPS also decreases overfitting problems through ensemble learning.

3) **PW transformation method**

Hüllermeier et al. proposed ranking by pairwise comparison (RPC) algorithm [31] in 2008 and applied it to multi-label classification. This algorithm transforms a multi-label problem which contains q labels into $q(q-1)/2$ binary classification subproblems and each subproblem is corresponding to one pair of labels. The subproblem of the label pair (y_i, y_j) contains all instances related with the label y_i or y_j in the original problem, but the instances which are related to these two labels simultaneously are excluded. In this way, instances related to the label y_i are positive instances, while the rest are viewed as negative instances. Hence, this subproblem can be solved by the traditional single-label algorithm. For a test example, all label pair patterns need to be tested, and then all labels can be sorted by the number of votes. The top labels have higher credibility.

Obviously, the scale of the RPC algorithm is influenced greatly by the number of labels. When there is a considerable volume of labels, the RPC algorithm is impractical for the high complexity. Moreover, the RPC algorithm fails to distinguish labels related to the test instance. In other words, it lacks a threshold or a division point to distinguish which part of labels belong to the instance.

Fürnkranz et al [26] proposed the calibrated label ranking (CLR) algorithm to solve the abovementioned problem of division point. In the CLR algorithm, one calibrated label y_0 was added as the boundary between relevant labels and irrelevant labels. During training, compared to the RPC algorithm, each label y_j only has to add one subproblem (y_j, y_0), in which the data cover all instances related to the label y_j (viewed as irrelevant with y_0) and instances unrelated with the label y_j (viewed as relevant to y_0).

2.2. **Algorithm Transformation Methods.** Algorithm transformation means to revise existing single-label algorithms to apply to multi-label data easily. Almost all single-label algorithms can be adapted to the learning from multi-label data. In this section, several representatives and extensively used algorithm transformation methods were introduced, including the nearest neighbor, decision-making tree, neural network, and support vector machine.

1) **Nearest neighbor**

ML-kNN [113] is the first multi-label algorithm that uses the nearest neighbor technology. The basic idea of ML-kNN is to count the occurrence of labels in the k nearest neighbors and then compute the probability for each label under different times of occurrence. Prediction results were given according to the maximum posterior principle. To estimate the corresponding probability, ML-kNN has to implement abundant calculations and comparisons of distances. As a result, the time complexity is relatively high. In addition, the effect of ML-kNN algorithm is easy to be affected if there is noise in the training set. This is the common fault of all algorithms based on the nearest neighbor technology. And, ML-kNN does not consider label correlations. To address this problem, LPLC [37] utilizes the correlations in a pair of labels within the nearest neighbor range. LPLC shows small differences in probability estimation from ML-kNN, and key difference lies in that LPLC devote to find relevant label set for the predicted labels. LPLC assumes that the strong correlations only exist among labels related to training instances. For a training set with n instances and q labels, it is necessary to define a $n \times q$ matrix M to record relevant labels related with each instance. Then, the probability for each label is computed based on M.

2) **Decision-making tree algorithm**

Clare and King proposed a multi-label algorithm ML-C4.5 [73] based on the decision-making tree C4.5 [23]. It constructs a decision-making tree from upper to bottom and the tree root includes all training samples. For instances in a non-leaf node, each characteristic

is investigated one by one to find the appropriate division point. This division point was used to partition the instances of this node, thus getting the maximum information gain. Clus-HMC [90] implemented hierarchical multi-label classification learning based on the predictive clustering trees (PCT) [10]. Similar to other decision-making trees, PCT also divides the current cluster into smaller clusters from top to bottom according to the principle of reducing intra-cluster variance to the maximum extent. For instances in the current cluster (S), the variance degree is defined as the sum of square of distance between label vector ($\mathbf{c}_i$) and the mean label vector ($\overline{\mathbf{c}}$).

3) **Neural network**

Bp-MLL [114] is the first algorithm that transforms the traditional neural network to apply to multi-label classification. It constructs a simple three-layer network, including the input layer, hidden layer, and output layer. The input layer contains d input units and each unit corresponds to a feature of the training set. The hidden layer contains M units. The output layer has q units and each unit corresponds to a label. The full connection is applied from the input layer to the hidden layer and from the hidden layer to the output layer. In Bp-MLL, the global loss function engages in distinguishing relevant labels and irrelevant labels of the instances and guides the learning system to output a relatively large value for relevant labels as well as a relatively small value for irrelevant labels. Compared to the traditional loss evaluating method that compares predicting values of each label in the output layer and real values directly, the Bp-MLL considers relations among different labels, thus achieving a better effect. CA2E [104] is an algorithm that is proposed recently for multi-label classification based on the deep neural network (DNN). The objective function of the CA2E algorithm can be divided into two parts. The first part is to solve the objective function using the DNN model [3] to get the embedded feature and label space. The second part aims to make the output of the whole model to recover the label space, in which the method similar to Bp-MLL is used to solve the problem.

4) **Support vector machine**

A. Elisseeff and J. Weston proposed the Ranking-SVM algorithm [25] in 2001. Firstly, the Ranking-SVM algorithm transforms the support vector machines (SVM) [11] which is a traditional high-efficiency single-label approach into the method directly used for multi-label classification. Ranking-SVM firstly defines the linear classifier $\{h_j(X) = \langle w_j, X\rangle + b_j = w_j^T X + b_j | 1 \leq j \leq q\}$ for each label and then maximizes the distances between all relevant-irrelevant label pairs. The ranking-SVM algorithm was improved in Reference [45].

3. **Challenges and Opportunities on Multi-Label Classification.** Multi-label classification is originated from text classification and it has become one of the most widely studied machine learning frameworks. Learning problems in many fields have to be solved by multi-label classification. At present, data acquisition technology is developed [103, 98] and multi-label data in many application fields is characteristic of large size as well as abundant features and labels. To cope with those data and further increase the classification effect, current researches on multi-label classification have new trends, which mainly concentrate on the utility of label correlation, label specific features, and feature compression [100]. These problems are the primary impetus to promote the performance of multi-label classification. In this section, these research contents are summarized and discussed thoroughly.

3.1. **Main Challenges.** In multi-label problems, labels often occur with interdependence rather than independence. Therefore, the key problem of multi-label classification lies in

the use of label correlations. For example, if one picture has the label "desert", it might have the label "camel" rather than "tree". In other words, "desert" and "camel" have very strong dependence. In contrast, the mutual dependence between "desert" and "tree" is significantly weaker. Many studies have proved that making use of label correlations can increase the performance of multi-label classification effectively [115, 101, 32, 44, 68, 46]. Due to the high dimensions and sparsity of label space, how to use label correlations is a quite complicated problem.

Compared with single-label problems, multi-label data often has large feature spaces. To describe semantics expressed by multi-labels, the number of features even can reach tens of thousands [52, 69, 62]. Some of these features are redundant or irrelevant for the classification tasks. Moreover, high-dimensional feature space often brings negative impacts on classification tasks, such as computing burden, overfitting, and decreasing classification effect [97, 41]. Therefore, many types of research concentrate on feature compression technique to get a low-dimension expression of multi-label data, thus constructing an algorithm with stronger generalization ability to improve the classification effect. Most feature compression algorithms construct an identical low-dimensional feature space for all labels [116, 2, 43]. In other words, different labels share the same feature expression. In multi-label problems, different labels reflect different semantics. Therefore, each label shall have corresponding unique features. These features are mostly related to corresponding labels and they are the most appropriate to distinguish labels [117, 42]. For example, in the task of image annotation, features related to textures are more appropriate to judge whether the image has the label "desert". Features related to colors are particularly appropriate to judge whether the image has the label "sky". In this study, such features are called "label-specific features". Zhang Minling et al. [117] proposed the concept of label-specific features for the first time in 2015 for multi-label classification and constructed the well-known LIFT algorithm. A lot of experiments on multiple benchmark datasets demonstrate that LIFT has an outstanding classification effect. Therefore, research methods based on label-specific features have attracted amounts of attention in recent years, and many improved algorithms like ML-DFL [118], WFSI-LIFT [107], LF-LPLC [100] and LLSF-DL [40] have been produced. Exploiting label-specific features brings breakthroughs to some practical challenges. For instance, in the classification task of protein subcellular localization modes based on biological images, extracting label-specific features can choose the appropriate feature subset from tens of thousands of protein features to describe the image [82], thus improving the classification effect.

3.2. **Label Correlations.**

1) **Types of label correlations**

In multi-label problems, labels do not occur independently but present some dependence. In other words, some labels often show up together in many instances, while other labels hardly co-occur. Using such a relationship is conducive to learn the more efficient and robust classification model [32]. In some problems, some labels have very few positive samples. Under this circumstance, it is extremely important to use label correlations. Making full use of label correlations has become one of the major research directions on multi-label classification at present and it is an important component in many algorithms [101, 32, 84, 24, 6]. In a word, existing using strategies of label correlations can be divided into three types [115]: First-order, second-order and high-order.

- First-order algorithms

First-order algorithms, including BR [12], ML-C4.5 [23] and ML-kNN [113], ignore label correlations completely. BR builds binary classification methods for each label. For

the label y_i, the objective is to learn the corresponding classifier $h_i : X \to \{0, 1\}$. In the process of learning, inputs are the original feature space, while outputs are values of the label y_i. Hence, different labels have the same inputs, but different outputs. For a multi-label classification problem involving q labels, the whole multi-label classifier $h : X \to \{0, 1\}^q$ is the combination of binary classifiers: $h = \{h_1, h_2, \cdots, h_q\}$. ML-C4.5 divides the training dataset into several small subsets layer by layer through a decision-making tree and the tree root covers all training data. For non-leaf nodes, some indexes like information entropy and Gini index are applied to further divide non-leaf nodes into sub-nodes, making the "purity" of data in sub-nodes higher than that of father nodes. ML-kNN is a lazy learning method and it does not learn the real model. At prediction, ML-kNN determines whether the test instance is related to a label by maximizing a posterior principle according to the distribution of labels in its k nearest neighbors in the training data.

- Second-order algorithms

The second-order algorithms, such as Rank-SVM [25], CLR [26], MLPP [66] and TSVA [67], considered pairwise label correlations. Rank-SVM defines the optimization goal of maximizing distance in a relevance-irrelevance label pair and solves the multi-label classification problem by SVM technology. CLR algorithm adds a virtual label y_0 into the original label set as the boundary point and builds a binary classification model for each label. Later, prediction results of all binary classification models can be ordered by voting. Labels before y_0 on the chain are relevant labels and the rest labels are irrelevant labels. MLPP trains classifiers by each pair of labels and determines the sequence of label relevance by combining prediction results of various classifiers through voting. TSVA is similar to CLR and it discusses several voting methods.

- High-order algorithms

High-order algorithms, such as RAkEL [85] and CC [74], consider correlations between several or all labels. As an ensemble learning method, RAkEL collects m LP classifiers, and the results of a test instance can be gained through voting among the results of all those models. Each classifier deals with a k-label subset and then transforms these subsets into a single-label classification problem. CC algorithm puts labels into one sequence randomly and then builds binary classifiers for each label in the sequence. For any label, labels in the front part of the sequence are added into the original features as new features. Relations among multiple labels are exploited in this way.

In terms of relation extraction or use perspective, high-order algorithms can be divided into global relation algorithms, local relation algorithm, and global-local combined relation algorithm.

- Global relation algorithm

The global relation algorithm believes that label correlations are global. In other words, the correlations between labels exist in all training data. CC [74], RPC [26], MLLS [46] and ML-LRC [102] belong to global relation algorithms.

- Local relation algorithm

In the local relation algorithm, label correlations exist in some training data. For example, "apple" and "fruit" have a very strong relation in gourmet magazines. However, "apple" and "digital equipment" often occur in technical journals together. Obviously, the label dependence relations only exist in some data in this case. If such label correlations are extracted or used from the global perspective, unnecessary and even misguided constraints will be imposed over all instances, which will lower the classification performance of models. These algorithms considering label correlations locally include LPLC [32],

ML-LOC [39], and so on. For example, ML-LOC divides training data into m groups, and the instances in a group share label correlation.

- Global-local combined relation algorithm

Global-local combined relation algorithms believe that considering both global and local label correlations simultaneously is conducive to establish a high-efficiency classification model. GLOCAL [119] is a global-local combined relation algorithm.

2) **Utilizing label correlations**

- Feature compression based on label correlations

To eliminate redundant and irrelevant features, researchers proposed many methods to compress multi-label data features. Many of these methods select features or implement feature transformation by using label correlations [42, 40, 82, 61, 49, 96]. For example, in Reference [61], a filtering feature selection method was proposed, which was called min-Redundancy and Max-Relevance (mRMR). To guide feature selection based on label correlations, the importance of each label was measured by mutual information. Subsequently, relations between features and labels are estimated by multiple weight strategies. If mutual information between two variables is zero, these two variables are completely independent; otherwise, the relation intensity between two variables is positively related to the value of mutual information.

- Feature extension based on label correlations

The abovementioned methods compress the original feature spaces based on label correlations. Some algorithms extend features by using label correlations, such as stacking based BR algorithm based on pruning [84], in which binary classifiers in the second layer of BR choose strongly related outputs from the first layer to extend the original feature space. In Reference [59], a classifier chain model was proposed to deal with multi-label problems. The main innovation is that it chooses a directed acyclic graph for modeling the label correlations and measures label correlations by conditional entropy, thus maximizing the sum of dependences among all labels expressed in the graph. Later, the original feature space is extended according to this directed acyclic graph and prediction, i.e., the results of binary classifiers corresponding to the ancient labels are added in. Later, binary classifiers are trained based on extended features. ML-LOC [38] also is an algorithm of feature extension based on label correlations.

- Embedding label correlations into classifiers

In addition, many algorithms use label correlations directly into the constructed classification model. In these algorithms, statistics on the frequency of label concurrence is the most direct way to estimate label dependence. If two labels often are annotated with the same samples, they show strong dependence. For each label of any instance, LPLC [37] records another label that has the maximum frequency of concurrence with it in the k nearest neighbors as the relevant label. For a test sample, k nearest neighbors are found firstly from the training set. Then, whether the test sample is related to a label is determined by the probability of occurrence of relevant labels in these neighbors.

RELIAB [115] builds a classification model based on label correlations from a rarely concerned perspective. It believes that although each sample is often related to multiple labels in multi-label data, the relation degree varies. For this reason, the RELIAB algorithm proposed the concept of relative labeling-importance (LRI). Moreover, it defines LRI as the polynomial distribution on the label space and gets the target solution through label propagation, thus taking advantage of label correlations. Later, classifiers were constructed by using the maximum entropy model based on LRI values.

Huang Shengjun et al. proposed the MAHR algorithm [36]. The basic idea of the MAHR algorithm is that if two labels are mutually dependent, the hypothesis of one label is conducive to establish the hypothesis of another label. MAHR realizes the hypothesized reuse mechanism through the traditional boosting technique and trains the corresponding boosting learning machines for each label. In each boosting process, one label not only produces basic learning machines from its hypothesis space but also generates hypotheses as much as possible by using other labels. These hypotheses increase weights of strongly related labels in the learning process through weighting and integration but decrease weights of weakly related labels. As a result, the MAHR algorithm needs no prior knowledge or statistics of label concurrence to calculate the dependence degree of labels, but it finds and uses dependences of labels through boosting, thus improving the effects of classifiers.

In a word, although many methods have studied label relation from different perspectives, it is a great challenge at present. The label relation is a relatively fuzzy concept and it has to be defined in specific scenarios. Sometimes, label correlations can be defined based on field knowledge. For example, the functional structure of genes and proteins present an obvious hierarchical structure [14]. Such hierarchical structural relations among different labels can be learned through hierarchical clustering [70] or the Bayes network structure [5]. Generally speaking, most applications do not have such an obvious hierarchical structure. Therefore, the majority of algorithms must evaluate label correlations from training data. For example, label correlations are estimated by the concurrence frequency of labels [71] or by building the label kernel method [116]. These methods might bring problems like overfitting and the evaluation method of label concurrence is also easy to be affected by class imbalance [32].

3.3. **Feature Compression.** The goal of feature compression is to eliminate noises and redundant features from the original high-dimensional features, thus increasing the generalization ability of the learning model and lowering computing complexity. Feature compression techniques mainly cover two types [109, 60]: feature extraction and feature selection. The feature extraction technique transforms a high-dimensional feature space into a low-dimensional feature space through a transformation matrix. Under this circumstance, the low-dimensional features are completely different from the original features and their relations are difficult to be comprehended. The feature selection technique selects some features directly from the original features and eliminates redundant or irrelevant features, thus maintaining the meanings of the original features. Feature selection can be further divided into supervised compression, unsupervised compression, and semi-supervision compression according to the use of label information in the process of feature compression. Generally, supervised compression is used widely, because it can improve the effect of feature compression by using distinctive information on labels. However, it is difficult to get mass training data with complete labels in some applications. Hence, the supervised feature compression method is inevitable to be trapped in overfitting [108]. Unsupervised feature selection method often makes uses of statistical features of data [77], spatial structural information of instances [35], and coefficient constraint [80]. The key of semi-supervised feature selection algorithm likes in full use of labeled data and unlabeled data to improve the learning effect [20].

In terms of interactive relations with model training, feature compression can be divided into filter type, wrapper type, and embedding type. In the filter type, feature compression is separated with model training [58, 76, 57]. Among filter-type feature compression methods, qualities of features are evaluated according to ReliefF [76], mutual information [56], and information gain [55], thus deleting features with relatively low evaluation

values. Wrapper-type feature compression methods evaluate the importance of feature subset according to the effect of the multi-label classifier. Embedding-type feature compression methods integrate feature selection with classification [54, 121]. Of course, there are algorithms with feature compression combined above types of methods. For example, MNLB [122] firstly utilizes principal component analysis (PCA) for feature selection, and then furtherly eliminates redundant or irrelevant features by using the genetic algorithm. Obviously, feature compression of the MNLB algorithm contains the filter-type and wrapper-type methods.

To sum up, feature selection tasks in multi-label problems are more complicated than those in single-label problems. The fundamental reason is due to the significantly rich semantics of multi-label problems, i.e., there are generally more features and the complicated label correlations have to be considered.

3.4. **Label-Specific Features.** In 2015, Zhang Minling et al. proposed a LIFT algorithm [117], in which the concept of label-specific features was proposed for the first time. Most existing multi-label methods often use the identical instance expression to build classification models for different labels. In other words, different labels use the same feature matrix in the learning process. Nevertheless, the LIFT algorithm believes that labels shall have their unique expressions. For instance, in applications of automatic image annotation, features related to colors are appropriate to distinguish images with "sky" scenes from images without "sky" scenes. Moreover, features related to texture are appropriate to distinguish whether images are related to "desert". Therefore, different labels shall use the appropriate feature expression. The effect of the classification model can only be further improved by using these features. These features are called "label-specific features". The concept of label-specific features has evident differences from the concept of traditional feature compression. In the concept of traditional feature compression, generally provides a unified feature expression to all labels through feature extraction or feature selection. Label-specific features refer to features that are related to specific labels rather than all labels.

At present, there are two main methods to construct label-specific features. One is feature extraction and the other is feature selection. The former one is represented by LIFT [117], while the latter one is represented by LLSF [42].

1) **Feature extraction based label-specific features**

LIFT extracts label-specific features for each label through feature extraction. Specifically, samples related to any label are viewed as positive samples, while the rest samples are viewed as negative samples. The k-means [48] are utilized to cluster over the positive sample set and negative sample set, respectively. Next, distances from each original sample to the center of these clusters are calculated, which form the new sample features. Next, binary classifiers are learned on spaces of these new label-specific features. For different labels, distributions of positive and negative samples are different, so that the built label-specific features vary from each other. Based on abundant experiments, the extraordinary effect of LIFT has been proved. Since then, label-specific features have attracted extensive interest from the academic circle and a series of algorithms have been proposed successively.

Based on LIFT, ML-DFL [118] discovers local structural relations between positive and negative sample sets through spectral clustering, thus transforming the original features into more distinctive features. WFSI-LIFT [107] thought that LIFT has ignored relations among samples and simply views that the importance of each sample is equal to each label, thus resulting in redundant features in label-specific features. Therefore, it allocates different weights for each sample according to the imbalance rate and then

extracts label-specific features by using the information theory, achieving a good effect. Y. Guo et al. [30] think it is not suitable that LIFT appoints the same number of clusters to positive and negative sample sets for different labels. Moreover, LIFT constructs label-specific features based on distance only, and it cannot make full use of relevant information from clustering results. To overcome these disadvantages, they proposed the LSDM algorithm which is composed of class information extraction, distance mapping feature construction, linear expression feature construction, linear discriminant analysis, and classification model generation. In class information extraction, the labeled positive sample set and negative sample set were mainly clustered by the clustering algorithm. For any label, the positive sample set and negative sample set have the same number of clusters. However, different labels have different numbers of clusters, which is different from LIFT. Later, label-specific features are extracted in distance mapping feature construction and linear expression feature construction by combining distance and linear methods. The linear discriminant analysis determines the number of clusters for different labels by using the optimization method, which avoids exhaustive search. After label-specific features are extracted, binary classification models are constructed based on these features. X. Wei et al. [95] believed that LIFT is inevitable to be trapped in local optimization by using the k-means clustering algorithm, and the clustering algorithm of high-dimensional data is very unstable. Hence, they extended the LIFT algorithm based on the idea of an integrated algorithm and proposed the ELIFT algorithm. In ELIFT, each label acquires several subsets from training sets by bagging technique and then trains binary classification models for each subset. The weighted sum of these binary classification models forms the overall classification model of the label. LF-LPLC [100] integrates label-specific features and local pairwise label correlation simultaneously, in which the specific features of each label are expanded by uniting the related features from the correlated labels. With such a framework, it enriches the labels' semantic information and solves the class-imbalanced problem.

2) **Feature selection based label-specific features**

The above algorithms all adopt feature transformation in the extraction of label-specific features. However, the LLSF algorithm proposed by J. Huang et al. [42] extracts label-specific features through feature selection technique. LLSF assumes that each label is only related to some of the original features and it expresses such sparsity in a linear regression with ℓ_1 constraints. In this way, the calculated linear regression parameters are not zero, indicating that the corresponding features are label-specific features; otherwise, the features are not label-specific features.

The optimization objective function defined by the LLSF algorithm also assures that there exist more label-specific features among highly similar labels, but few label-specific features among lowly similar labels. Since the LLSF algorithm implements feature selection through linear regression, it can learn binary classification models based on the selected features. On the other hand, the LLSF algorithm can predict models directly by the product of linear regression parameter vectors and feature vectors of test samples. NSLSF [99] considered in some applications the sparsity assumption does not hold, and proposed a novel feature selection-based approach for extracting label specific features. It translates the logic labels to the numerical ones to convey more semantic information and embed the label correlations. Then, a linear regression is modeled to describe the discrimination of label specific features based on the numerical labels.

Although the concept of label-specific features is proposed recently, it has attracted extensive attention between scholars. Existing associated models often can get effects from different problem fields and show good generalization by the classification model

which is constructed based on label-specific features. Label-specific features belong to specific labels. Hence, it is convenient to construct a binary classification model by using the traditional and mature single-label classification algorithm, which is very applicable to solve practical problems. Current studies on label-specific features often are combined with label correlations and sample relations. Many studies have discussed label-specific features from perspectives of label missing and class imbalance.

3.5. **Other research directions.**

1) **Class imbalance**

A database with millions of pictures might involve tens of thousands of scene labels, but the quantity of images with the label "grassland" is significantly lower compared with the number of images without the label "grassland". Therefore, multi-label problems are often class imbalanced. Such class imbalance is more significant in the situation with a great number of labels [120]. The class imbalance problem is viewed as a basic challenge in multi-label learning, thus lowering the effects of most multi-label learning methods [112, 34].

The class imbalance problems also exist in traditional binary-class and multi-class problems universally. Hence, the solving idea can provide some enlightenments to process multi-label imbalance problems. Solving methods to these traditional class imbalance problems are generally divided into two types: data preprocessing [47] and cost-sensitive classification [27]. Data preprocessing includes undersampling that deletes majority class data and oversampling that adds minority data. Algorithms like RUS [50], ROS, and SMOTE [19] all belong to data preprocessing. After data preprocessing, algorithms that are appropriate for problems can be constructed on the data. Cost sensitivity classification generally gives a cost penalty for the wrong classification in the construction of a classification model, thus optimizing the learning object toward the reduction of cost, such as CS-LDM [18] and boosting weight ELM [53].

M. -L. Zhang et al. proposed a COCOA algorithm [123] to solve the problem of class imbalance in multi-label applications. In COCOA, for each label y_j, the classification model f_j is formed by linear addition of one binary classification model and K multi-class classification models. Moreover, the threshold constant (a_j) is determined by maximizing the corresponding F-measure value. When $f_j(\mathbf{x}) > a_j$, $\mathbf{x}$ is a positive sample; otherwise, it is a negative sample. The construction of a binary classification model is similar to BR. For the multi-class classification models, COCOR selects the subset (excluding y_j) containing k labels randomly and labels in this subset form k label pairs with y_j. Then, the k multi-class classification models are transformed into a multi-class problem according to the values of each label pair and each multi-class problem learns one classification model. Different from the LP transformation technique, COCOR combines classes with small quantities in the transformation of a multi-class problem, thus solving the class imbalance.

F. Charte et al. [17] proposed several complicated measurement standards for multi-label imbalance degrees, including IRperLabel, MeanIR, and CVIR. Meanwhile, F. Charte et al. also proposed undersampling and oversampling algorithms (LP-RUS and LP-ROS) for preprocessing of multi-label data imbalance. Models based on LP transformation [12] are constructed and prove that these sampling techniques are valid. The MLSMOTE algorithm [16] discussed an oversampling technique for multi-label learning which used a similar strategy with SMOTE [19]. Firstly, the MLSMOTE algorithm recognizes the labels of minority through MeanIR and CVIR and then synthesizes new samples for these labels. Similar to the method proposed by the SMOTE algorithm, features and labels of these new samples are gained through interpolation of nearest samples. In this study,

three synthesis strategies are discussed, including intersection, sum, and sequence-based selection.

In a word, common and intuitive processing of class unbalance in multi-label problems is realized by sampling. Further studies on how to use label correlations in sampling are needed. Additionally, adverse impacts of class unbalance can be eliminated as much as possible by ensemble learning and cost sensitivity. However, there are a few associated studies.

2) **Label missing**

In many real-world applications, it is impractical to gain all real relevant labels of all samples in the training set for the following two main reasons [9, 94]. On one hand, many applications contain a lot of label classes that require considerable labor forces. On the other hand, meanings of different labels might overlap and it is difficult to be distinguished completely. Therefore, learning models based on data of such partial labels might be unable to capture label correlations and relations between labels and features accurately.

If unknown labels are recognized clearly, the problem becomes simpler and it only requires construction models based on known labels [106]. In most cases, the positions of the missing label are unknown. Under this circumstance, L. Xu et al. [102] proposed a multi-label learning framework. This framework not only can learn label correlations automatically in the process of model learning and capture the high-order complicated relations of labels through a lower-ranking structure but also supplements label matrix based on such high-order complicated relations, thus solving the problem of label missing.

3) **Label compression**

Since the number of labels in many practical multi-label problems can reach tens of thousands, many studies have shift attention to the multi-label classification which involves abundant labels. Excessive labels might bring considerable time and space costs to the algorithm. Moreover, many mature and common algorithms like BR [12] and ECC [74] are inapplicable to handle these excessive labels. To address this problem, researchers proposed the spatial dimension compression technique for labels. Firstly, high-dimensional labels are compressed into a low-dimensional label space, and the classification models are trained in the low-dimensional label space to decrease computing burdens. Of course, prediction results in the low-dimensional label space have to be recovered into the original feature space. According to existing studies, label compression not only can shorten the operation time of the algorithm but also can improve the classification effect [87].

Existing label compression algorithms basically can be divided into two types [86]: label transformation and label subset. Label compression algorithms based on label transformation generally take advantage of projection mechanisms, such as compressed sensing [33], PCA [87], and so on. To sum up, label compression algorithms based on label transformation have many advantages, such as a strong theoretical basis, easy implementation, etc. However, labels after transformation lack of meaning of original labels and they are difficult to be connected mutually. Label compression algorithms based on label subsets often use group sparse learning [8], random sampling [7], Boolean matrix decomposition [93], etc. These algorithms acquire low-dimensional labels, or labels are from the original labels directly, or the original feature space can be recovered completely. Therefore, the label compression algorithms based on label subsets have strong explanatory powers.

4. **Conclusions.** Firstly, this study illustrates an enormous variety of multi-label algorithms, including the detailed discussion of the representative algorithms and their variants. Secondly, main research directions including label-specific features and label

correlations, representative algorithm, and improvement processes are introduced with consideration to the main challenges of multi-label algorithms in recent years. Hence, this study provides a comprehensive perspective to understand research status and development directions of multi-label classification problems in recent years, thus enabling to promote further progress of relevant studies. This study involves the recent challenges of multi-label classification, and presents the emerging trends to deal with them. It gives us some opportunities to employ multi-label learning for a variety of applications in specific fields. The combination of multi-label classification and other machine learning tasks is another research direction that stimulate new methods. For example, the combination of multi-label classification and recommendation can provide new basic data for a recommendation, thus enabling to produce of a more effective model.

Acknowledgment. This research was supported by the Natural Science Foundation of Jiangsu Province of China (BK20200364), the Natural Science Foundation of the Fujian Province of China (2018J01572), and the National Natural Science Foundation of China (62001111).

REFERENCES

[1] R. Agrawal, A. Gupta, Y. Prabhu, M. Varma, Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages, *Proc. of the 22nd international conference on World Wide Web*, pp.13–24, 2013.

[2] A. Alalga, K. Benabdeslem, N. Taleb, Soft-constrained Laplacian score for semi-supervised multi-label feature selection, *Knowl. Inf. Syst.*, vol. 47, no. 1, pp.75–98, 2015.

[3] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in *Proc. of The 30th International Conference on Machine Learning*, pp.1247–1255, 2013.

[4] Z. Barutcuoglu, R. Schapire, O. Troyanskaya, Hierarchical multi-label prediction of gene function, *Bioinformatics*, vol. 22, no. 7, pp.830–836, 2006.

[5] W. Bi, J. T. Kwok, Multilabel classification with label correlations and missing labels, *Proc. of 28th AAAI Conference on Artificial Intelligence*, pp.1680–1686, 2014.

[6] N.C. Bianchi, C. Gentile, L. Zaniboni, Incremental algorithms for hierarchical classification, *Journal of Machine Learning Research*, vol. 7, pp.31-54, 2006.

[7] W. Bi, J. T. Y. Kwok, Efficient Multi-label Classification with Many Labels, in *international conference on machine learning*, pp.405–413, 2013.

[8] K. Balasubramanian, G. Lebanon, The Landmark Selection Method for Multiple Output Prediction, in *29th international conference on machine learning*, 2012.

[9] S. Bucak, R. Jin, A. Jain, Multi-label learning with incomplete class assignments, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2801–2808, 2011.

[10] H. Blockeel, L. De Raedt, J. Ramon, Top-down induction of clustering trees, in *Proc. of the 15th Intel Conf. on Machine Learning*, pp.55–63, 1998.

[11] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in *the Fifth Annual Workshop on Computational Learning Theory*, pp.144–152, 1992.

[12] M. R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning multi-label scene classification, *Pattern Recognition*, vol. 37, pp.1757–1771, 2004.

[13] K. Bhatia, H. Jain, P. Kar, M. Varma, P. Jain, Sparse local embeddings for extreme multi-label classification, *Proc. of the 28th International Conference on Neural Information Processing Systems*, vol. 1, pp.730–738, 2015.

[14] R. Cerri, R. C. Barros, P. L. F. de Carvalho A. C. , Y. Jin, Reduction strategies for hierarchical multi-label classification in protein function prediction, *BMC Bioinformatics*, vol. 17, no. 1, pp.1–24, 2016.

[15] G. Carneiro, A. Chan, P. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp.294–410, 2007.

[16] F. Charte, A. J. Rivera, M. J. Del Jess, F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, *Knowledge-Based Systems*, vol. 89, pp.385–397, 2015.

[17] F. Charte, A. Rivera, M. J. Del Jesus, F. Herrera, A First Approach to Deal with Imbalance in Multi-label Datasets, in *8th International Conference on Hybrid Artificial Intelligent Systems*, pp.150–160, 2013.

[18] F. Cheng, J. Zhang, C. Wen, Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data, *Pattern Recognition Letters*, vol. 80, pp.107–112, 2016.

[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, vol. 16, pp.321–357, 2002.

[20] X. Chang, F. Nie, Y. Yang, H. Huang, A convex formulation for semi-supervised multi-label feature selection, *Proc. of the 28th AAAI Conference on Artificial Intelligence*, pp.1171–1177, 2014.

[21] Z. Chen, Z. Hao, A unified multi-label classification framework with supervised low-dimensional embedding, *Neurocomputing*, vol. 171, pp.1563–1575, 2016.

[22] E. A. Cherman, J. Metz, M. C. Monard, Incorporating label dependency into the binary relevance framework for multilabel classification, *Expert Syst Appl.*, vol. 39, no. 2, pp.1647–1655, 2012.

[23] A. Clare, R. D. King, Knowledge discovery in multi-label phenotype data, in *Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, Freiburg, Germany, pp.42–53, 2001.

[24] k. Dembczyński, w. -w. Cheng, E. Hüllermeier, Bayes optimal multilabel classification via probabilistic classifier chains, in *IEEE International Conference on Data Mining*, pp.1609–1614, 2010.

[25] A. Elisseeff, J. Westom, A kernel method for multi-labeled classification, in *Advances in Neural Information Processing Systems 14*, MIT Press, pp.681–687, 2001.

[26] J. Fürnkranz, E. Hüllermeier, E. L. Mencia, K. Brinker, Multilabel classification via calibrated label ranking, *Machine Learning*, vol. 73, no. 2, pp.133–153, 2008.

[27] F. Provost, T. Fawcett, Robust classification for imprecise environments, *Machine Learning*, vol. 42, pp.203–231, 2001.

[28] S. Godbole, S. Sarawagi, Discriminative methods for multi labeled classification, *Proc. of Pacific-Asia conference on knowledge discovery and data mining*, Springer Berlin Heidelberg, pp.22–30, 2004.

[29] E. C. Goncalves, A. Plastino, A. A. Freitas, Simpler Is Better: A Novel Genetic Algorithm to Induce Compact Multi-Label Chain Classifiers, *Proc. of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pp.559–566, 2015.

[30] Y. Guo, F. Chung, G. Li, J. Wang, J. C. Gee, Leveraging Label-Specific Discriminant Mapping Features for Multi-Label Learning, *ACM Transactions on Knowledge Discovery From Data*, vol. 13, no. 2, Article 24, 2019.

[31] E. Hüllermeier, J. F*ddot*urnkranz, W. W. Cheng, K. Brinker, Label Ranking by Learning Pairwise Preferences, *Artificial Intelligence*, vol. 172, pp.1897–1916, 2008.

[32] J. Huang, G. R. Li, S. H. Wang, Z. Xue, Q. M. Huang, Multi-Label Classification by Exploiting Local Positive and Negative Pairwise Label Correlation, *Neurocomputing*, vol. 257, pp.164–174, 2017.

[33] D. J. Hsu, S. Kakade, J. Langford, T. Zhang, Multi-label prediction via compressed sensing, *Computer Science*, vol. 22, pp.772–780, 2009.

[34] H. He, E. A. Garcia, Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp.1263–1284, 2009.

[35] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, *Advances in Neural Information Processing Systems*, pp.507–514, 2006.

[36] S.-J. Huang, Y. Yu, Z.-H. Zhou, Multi-label hypothesis reuse, *Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.525–533, 2012.

[37] J. Huang, G. Li, S. Wang, Q. Huang, Categorizing Social Multimedia by Neighborhood Decision Using Local Pairwise Label Correlation, in *IEEE International Conference on Data Mining Workshop*, pp.913–920, 2014.

[38] Q. Hu, W. Pedrycz, D.Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, *IEEETrans.Syst.Man Cybern, Part B: Cybern*, vol. 40, pp.137–150, 2010.

[39] S.-J. Huang, Z.-H. Zhou, Multi-label learning by exploiting label correlations locally, in *Proc. of the 26th AAAI Conference on Artificial Intelligence*, pp.949–955, 2012.

[40] J. Huang, G. Li, Q. Huang, X. Wu, Learning Label-Specific Features and Class-Dependent Labels for Multi-Label Classification, *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp.3309–3323, 2016.

[41] Q. Hu, W. Pan, L. Zhang, D. Zhang, Y. Song, D. Yu, Feature selection for monotonic classification, *IEEETrans.FuzzySyst.*, vol. 20, pp.69–81, 2012.

[42] J. Huang, G. R. Li, Q. M. Huang, X. D. Wu, Learning Label Specific Features for Multi-label Classification, in *2015 IEEE International Conference on Data Mining*, pp.181–190, 2015.
[43] R. Huang, W. Jiang, G. Sun, Manifold-based constraint Laplacian score for multi-label feature selection, *Pattern Recognition Letters*, vol. 112, pp.346–352, 2018.
[44] Z. F. He, M. Yang, Y. Gao, H. D. Liu, Y. Yin, Joint multi-label classification and label correlations with missing labels and feature selection, *Knowledge-Based Systems*, vol. 163, pp.145–158, 2019.
[45] A. W. Jiang, C. H. Wang, Y. P. Zhu, Calibrated Rank-SVM for multi-label image categorization, *IEEE International Joint Conference on. Neural Networks*, pp.1450–1455, 2008.
[46] S. Ji, L. Tang, S. Yu, J. Ye, Extracting shared subspace for multi-label classification, in *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pp.381–389, 2008.
[47] N. Japkowicz, *Learning from imbalanced data sets: A comparison of various strategies*, AAAI Press, 2000, pp.10–15.
[48] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review, *ACM Computing Surveys*, vol. 31, no. 3, pp.264–323, 1999.
[49] S. Jungjit, M. Michaelis, A. A. Freitas, J. Cinatl, Two Extensions to Multi-label Correlation-Based Feature Selection: A Case Study in Bioinformatics, in *IEEE International Conference on Systems, Man, and Cybernetics*, pp.1519–1524, 2013.
[50] S.B. Kotsiantis, P.E. Pinellas, Mixture of expert agents for handling imbalanced data sets, *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, pp.46–55, 2003.
[51] Y. Luo, T. Liu, D. Tao, C. Xu, Multiview Matrix Completion for Multilabel Image Classification, *IEEE Transactions on Image Processing*, vol. 24, no.8, pp.2355–2368, 2015.
[52] J. Lin, X. Sun, P. Yang, S. Ma, Q. Su, Semantic-Unit-Based Dilated Convolution for Multi-Label Text Classification, *Empirical Methods in Natural Language Processing*, pp.4554–4564, 2018.
[53] K. Li, X. Kong, Z. Lu, L. Wenyin, J. Yin, Boosting weighted ELM for imbalanced learning, *Neurocomputing*, vol. 128, pp.15–21, 2014.
[54] G.-Z. Li, M. You, L. Ge, J.Y. Yang, M. Q. Yang, Feature selection for semi-supervised multi-abel learning with application to gene function analysis, *Proc. of the First ACM International Conference on Bioinformatics and Computational Biology*, pp.354–357, 2010.
[55] L. Li, H. Liu, Z. Ma, Y. Mo, Z. Duan, J. Zhou, J. Zhao, Multi-label feature selection via information gain, in *International Conference on Advanced Data Mining and Applications*, pp.345–355, 2014.
[56] J. Lee, D.-W. Kim, Mutual information-based multi-label feature selection using interaction information, *Expert Syst. Appl.*, vol. 42, pp.2013–2025, 2015.
[57] F. Li, D. Miao, W. Pedrycz, Granular multi-label feature selection based on mutual information, *Pattern Recognition*, vol. 67, pp.410–423, 2017.
[58] J. Lee, D.-W. Kim, Feature selection for multi-label classification using multi-variate mutual information, *Pattern Recognit. Lett.*, vol. 34, pp.349–357, 2013.
[59] J. Lee, H. Kim, N. Kim, J.-H. Lee, An approach for multi-label classification by directed acyclic graph with label correlation maximization, *Information Sciences*, 351, pp 101- 114, 2016.
[60] Y. J. Lin, Q. H. Hu, J. H. Liu, J. K. C, J. Duan, Multi-label feature selection based on neighborhood mutual information, *Applied soft computing*, vol. 38, pp.244–258, 2016.
[61] L. Liu, J. Zhang, P. Li, Y. Zhang, X. Hu, A Label Correlation Based Weighting Feature Selection Approach for Multi-label Data, in *International Conference on Web-Age Information Management*, pp.369–379, 2016.
[62] Y. Lin, Q. Hu, J. Liu J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, *Neurocomputing*, vol. 168, pp. 92–103, 2015.
[63] Y. Liu, D. Zhang, G. Lu, Region-based image retrieval with high-level semantics using decision tree learning, *Pattern Recognition*, vol. 41, no. 8, pp.2554–2570, 2008.
[64] Y. Luo, T. Liu, D. Tao C. Xu, Multiview Matrix Completion for Multilabel Image Classification, *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp.2355–2368, 2015.
[65] A. McCallum, Multi-label text classification with a mixture model trained by EM, in *working Notes of the AAAI workshop on Text Learning*, Orlando, FL, 1999.
[66] E. L. Mencía, J. Fürnkranz, Pairwise learning of multilabel classifications with perceptrons, in *Proc. of the International Joint Conference on Neural Networks, Part of the IEEEWorld Congress on Computational Intelligence*, Hong Kong, China, pp.2899–2906, 2008.
[67] G. Madjarov, D. Gjorgjevikj, T. Delev, Efficient two-stage voting architecture for pairwise multi-label classification, in *Proc. of the Advances in Artificial Intelligence*, Springer, pp.164–173, 2011.

[68] T. T. T. Nguyen, A. V. Luong, Q. V. H. Nguyen, A. W.-C. Liew, B. Stantic, Multi-label classification via label correlation and first-order feature dependence in a data stream, *Pattern Recognition*, vol. 90, pp.35–51, 2019.

[69] P. Naula, A. Airola, T. Salakoski, T. Pahikala, Multi-label learning under feature extraction budgets, *Pattern Recognit.Lett.*, vol. 40, pp.56–65, 2014.

[70] K. Punera, S. Rajan, J. Ghosh, Automatically learning document taxonomies for hierarchical classification, *Proc. of the 14th International Conference on WWW*, pp.1010–1011, 2005.

[71] J. Petterson, T. Caetano, Submodular multi-label learning, *Advances in Neural Information Processing Systems*, pp.1512–1520, 2011

[72] G. J. Qi, X. S. Hua, Y. Rui, J. Tan, T. Mei, H. J. Zhang, Correlative multi-label video annotation, *Proc. of the 15th International Conference on Multimedia*, Augsburg, Germany, pp.17–26, 2007.

[73] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California, 1993.

[74] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, *Machine Learning*, vol. 85, pp.333–359, 2011.

[75] J. Read, B. Pfahringer, G. Holmes, Multi-label classification using ensembles of pruned sets, in *Eighth IEEE International Conference on Data Mining*, pp:995–1000, 2008.

[76] O. Reyes, C. Morell, S. Ventura, Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context, *Neurocomputing*, vol. 161, pp.168–182, 2015.

[77] Y. Ren, G. Zhang, G. Yu, X. Li, Local and global structure-preserving based feature selection, *Neurocomputing*, vol. 89, pp.147–157, 2012.

[78] S. Shi, Y. Chen, M. Fang, W. Li, Shining, A Hierarchical Multi-label Propagation Algorithm for Overlapping Community Discovery in Social Network, in *11th Web Information System and Application Conference*, pp.113–118, 2014.

[79] R. E. Schapire, Y. Singer, Boostexter: A boosting-based system for text categorization, *Machine Learning*, vol. 39, pp.135–168, 2000.

[80] C. Shi, Q. Ruan, G. An, Sparse feature selection based on graph Laplacian for web image annotation, *Image Vis. Comput.*, vol. 32, pp.189–201, 2014.

[81] L. Sun, M. Kudo, K. Kimura, READER: Robust Semi-Supervised Multi-Label Dimension Reduction,*IEICE Transactions on Information and Systems*, vol. 10, pp.2597–2604, 2017.

[82] W. Shao, M. Liu, Y.-Y. Xu, H.-B. Shen, D. Zhang, An Organelle Correlation-Guided Feature Selection Approach for Classifying Multi-Label Subcellular Bio-Images, *IEEE ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp.828–838, 2018.

[83] P. N. da Silva, E. C. Goncalves, A. Plastino, A. A. Freitas, Distinct chains for different instances: An effective strategy for multi-label classifier chains, in *Machine Learning and Knowledge Discovery in Databases - European Conference*, pp.453–468, 2014.

[84] G. Tsoumakas, A. Dimou, E. Spyromitros, V. Mezaris, I. Kompatsiaris, I. Vlahavas, Correlation-Based Pruning of Stacked Binary Relevance Models for Multi-Label Learning, in *1st International Workshop on Learning from Multi-Label Data*, pp.101–116, 2009.

[85] G. Tsoumakas, I. Vlahavas, Random k-label sets An ensemble method for multilabel classification, in *European Conference on Machine Learning*, pp.406–417, 2007.

[86] L. Tang, L. Liu, J. Gan, An Overview of Label Space Dimension Reduction for Multi-Label Classification, *Proc. of the 2nd International Conference on Intelligent Information Processing*, 2017.

[87] F. Ta, H. T. Lin, Multilabel classification with principal label space transformation, *Neural Computation*, vol. 24, no. 9, pp.2508–2542, 2012.

[88] M. A. Tahir, J. Kittler, A. Bouridane, Multilabel classification using heterogeneous ensemble of multi-label classifiers, *Pattern Recognition Letters*, vol. 33, no. 5, pp.513–523, 2012.

[89] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp.1–13, 2007.

[90] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, H. Blockeel, Decision trees for hierarchical multi-label classification, *Machine Learning*, vol. 73, no. 2, pp.185–214, 2008.

[91] J. Weston, S. Bengio, N. Usunier, WSABIE: scaling up to large vocabulary image annotation, *Proc. of the Twenty-Second international joint conference on Artificial Intelligence*, vol. 3, pp.2764–2770, 2011.

[92] D. H. Wolpert, Original Contribution: Stacked Generalization, *Neural Networks*, vol. 5, no. 2, pp.241–259, 1992.

[93] J. Wicker, B. Pfahringer, S. Kramer, Multi-label classification using boolean matrix decomposition, *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp.179–186, 2012.

[94] B. Wu, F. Jia, W. Liu, B. Ghanem, S. Lyu, Multi-label Learning with Missing Labels Using Mixed Dependency Graphs, *International Journal of Computer Vision*, vol. 126, no. 8, pp.875–896, 2018.
[95] X. Wei, Z. Yu, C. Zhang, Q. Hu, Ensemble of Label Specific Features for Multi-Label Classification, in *2018 IEEE International Conference on Multimedia and Expo*, pp.1–6, 2018.
[96] X. Wang, R.-C. Chen, C. Hong, Z. Zeng, Z. Zhou, Semi-supervised multi-label feature selection via label correlation analysis with l1-norm graph embedding, *Image and Vision Computing*, vol. 63, pp.10–23, 2017.
[97] X. Wu, X. Zhu, G. Wu, W. Ding, Data mining with big data, *IEEE Trans. Knowl. Data Eng.*, vol. 26, pp.97–107, 2014.
[98] J. Wang, W. M. Song, X. y. Sun, L. L. Tang, J. -H Yeh, Annotation Method to Improve the Mapping Between Image Features and High Level Semantic Expression, *Journal of Network Intelligence*, Vol. 5, No. 4, pp. 211-217, 2020.
[99] W. Weng, Y.-N. Chen, C.-L. Chen, S.-X. Wu, and J.-H. Liu, Non-sparse Label Specific Features Selection for Multi-label Classification, *Neurocomputing*, vol. 377, pp.85–94, 2020.
[100] W. Weng, Y. Lin, S. Wu, Y. Li, Y. Kang, Multi-label learning based on label specific features and local pairwise label correlation, *Neurocomputing*, vol. 273, pp.385–394, 2018.
[101] J. Y. Xu, J. Ma, Correlation-Based Weighted K-Label sets for Multi-Label Classification, in *Asia-Pacific Web Conference*, pp.408–419, 2016.
[102] L. Xu, Z. Wang, Z. Shen, Y. Wang, E. Chen, Learning Low-Rank Label Correlations for Multi-label Classification with Missing Labels, in *IEEE International Conference on Data Mining*, pp.1067–1072, 2014.
[103] S. Yang, X. -f. Wang, Sparse Representation and SRCNN based Spatio-temporal Information Fusion Method of Multi-sensor Remote Sensing Data, *Journal of Network Intelligence*, Vol. 6, No. 1, pp. 40-53, 2021.
[104] C.-K. Yeh, W.-C. Wu, W.-J. Ko, Y.-C. F. Wang, Learning Deep Latent Spaces for Multi-Label Classification, *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence*, pp.2838–2844, 2017.
[105] A. Yüce, H. Gao, J.-P. Thiran, Discriminant multi-label manifold embedding for facial Action Unit detection, in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 6, pp.1–6, 2015.
[106] H. F. Yu, P. Jain, P. Kar, I. S. Dhillon, Large-scale Multi-label Learning with Missing Labels, *International Conference on Machine Learning*, pp.593–601, 2014.
[107] Y. Yan, S. Li, Z. Yang, X. Zhang, J. Li, A. Wang, J. Zhang, Multi-label Learning with Label-Specific Feature Selection, in *International Conference on Neural Information Processing*, pp.305–315, 2017.
[108] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, *Int. J. Comput. Vis.*, vol. 113, pp.113–127, 2015.
[109] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, A.G. Hauptmann, Multi-feature fusion via hierarchical regression for multimedia analysis, *IEEE Trans. Multimedia*, vol. 15, pp.572–581, 2013.
[110] T. Yu, W. Zhang, Semisupervised Multilabel Learning With Joint Dimensionality Reduction, *IEEE Signal Processing Letters*, vol. 23, no. 6, pp.795–799, 2016.
[111] D. Zha, C. Li, Multi-label dataless text classification with topic modeling, *Knowledge and Information Systems*, vol. 61, pp.137–160, 2019.
[112] M. L. Zhang Z. H. Zhou, A review on multi-label learning algorithms, *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp.1819–1837, 2014.
[113] M. L. Zhang, Z. H. Zhou, Ml-knn: A lazy learning approach to multi-label learning, *Pattern Recognition*, vol. 40, pp.2038–2048, 2007.
[114] M.-L. Zhang, Z.-H. Zhou, Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization, *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp.1338–1351, 2006.
[115] Y. K. Li, M. L. Zhang, X. Geng, Leveraging Implicit Relative Labeling-Importance Information for Effective Multi-label Learning, in *IEEE International Conference on Data Mining*, pp.251–260, 2015.
[116] Y. Zhang, Z.-H. Zhou, Multilabel dimensionality reduction via dependence maximization, *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 3, pp.1–21, 2010.
[117] M.-L. Zhang, L. Wu, Lift: Multi-Label Learning with Label-Specific Features, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp.107–120, 2015.
[118] J. J. Zhang, M. Fang, X. Li, Multi-label learning with discriminative features for each label, *Neurocomputing*, vol. 154, pp.305–316, 2014.

[119] Y. Zhu, J. T. Kwok, Z.-H. Zhou, Multi-Label Learning with Global and Local Label Correlation, *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp.1081–1094, 2018.

[120] Z.-H. Zhang, S.-N. Li, Z.-G. Li, H. Chen, Multi-label feature selection algorithm based on information entropy, *J. Comput. Res. Dev.*, vol. 50, no. 6, pp.1177–1184. (in Chinese).

[121] P. Zhu, Q. Xu, Q. Hu, C. Zhang, H. Zhao, Multi-label feature selection with missing labels, *Pattern Recognit.*, vol. 74, pp.488–502, 2018.

[122] M.-L. Zhang, J. M. Pena, V. Robles, Feature selection for multi-label naive Bayes classification, *Information Sciences*, vol. 179, no. 19, pp.3218–3229, 2009.

[123] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, Towards class-imbalance aware multi-label learning, in *Proc. of the 24th International Conference on Artificial Intelligence*, pp.4041–4047, 2015.