Published in final edited form as:
ACM Comput Surv. 2020 June ; 53(2): . doi:10.1145/3379504.

Multi-Label Active Learning Algorithms for Image Classification: Overview and Future Promise

JIAN WU,
Soochow University, China and Human Longevity, Inc., USA

VICTOR S. SHENG,
Texas Tech University, USA

JING ZHANG,
Nanjing University of Science and Technology, China

HUA LI,
Washington University in St. Louis, USA

TETIANA DADAKOVA,
Human Longevity, Inc., USA

CHRISTINE LEON SWISHER,
Human Longevity, Inc., USA

ZHIMING CUI,
Suzhou University of Science and Technology, China

PENGPENG ZHAO
Soochow University, China

Abstract

Image classification is a key task in image understanding, and multi-label image classification has become a popular topic in recent years. However, the success of multi-label image classification is closely related to the way of constructing a training set. As active learning aims to construct an effective training set through iteratively selecting the most informative examples to query labels from annotators, it was introduced into multi-label image classification. Accordingly, multi-label active learning is becoming an important research direction. In this work, we first

corresponding authors: victor.sheng@ttu.edu (V. S. Sheng), ppzhao@suda.edu.cn (P. Zhao).
Authors' addresses:
J. Wu, Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006, China, Human Longevity, Inc., San Diego, CA, 92121; V. S. Sheng, Department of Computer Science, Texas Tech University, Lubbock, TX, 79409; J. Zhang, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China; H. Li, Department of Radiation Oncology, Washington University in St. Louis, St. Louis, MO, 63110; T. Dadakova and C. L. Swisher, Human Longevity, Inc., San Diego, CA, 92121; Z. Cui, School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu, 215009, China; P. Zhao (corresponding author), Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006, China.

review existing multi-label active learning algorithms for image classification. These algorithms can be categorized into two top groups from two aspects respectively: sampling and annotation. The most important component of multi-label active learning is to design an effective sampling strategy that actively selects the examples with the highest informativeness from an unlabeled data pool, according to various information measures. Thus, different informativeness measures are emphasized in this survey. Furthermore, this work also makes a deep investigation on existing challenging issues and future promises in multi-label active learning with a focus on four core aspects: example dimension, label dimension, annotation, and application extension.

Keywords

Additional Key Words and Phrases; Image classification; multi-label image; active learning; sampling strategy; annotation

1 INTRODUCTION

In the era of big data, with the rapid advances in computer technology, networks, and the information industry, the amount of various types of data is increasing in an explosive way. According to the reports of International Data Corporation (IDC), the amount of data around the world will increase sharply from 0.8 ZB in 2009 to 35 ZB in 2020, which is a 44 times growth over 10 years with an average annual growth of 40% [61]. This new paradigm leads to an enormous challenge on the management of the data. Among various types of data, multimedia data accounts for the largest proportion. To make full use of the large-scale multimedia data, understanding the semantic information of multimedia data becomes a primary task. Consequently, feasible approaches for automatically understanding multimedia data are highly in demand [88].

Aiming to annotate images with specific labels for describing their semantic information, image classification has become a key task in image understanding. In image classification, high-level semantic information of images can be extracted from their underlying features, and predefined labels can be assigned to these images. More precisely, a classifier can be learned from a training set using a certain classification algorithm to classify unlabeled images. Therefore, image classification receives much attention and has been deemed an important issue in computer vision and pattern recognition communities [99, 117]. Image classification is widely applicable to many domains, including object recognition, scene recognition, image retrieval [11], video annotation [90], and so on [13, 37, 43, 81], in the real world and widely used in medical diagnosis, aerospace, military reconnaissance, fingerprint recognition [51], nondestructive inspection, remote sensing satellite, and other fields as well. Clearly, image classification plays an increasingly important role in industry and business.

Training a high-performance classification model is the core of image classification. The quality of a trained classification model heavily relies on its training dataset. Constructing an effective training set for image classification is an interesting and significant task. In the era of big data, a large number of unlabeled images is yielded. However, obtaining a large number of labeled images is time-consuming and needs a lot of manpower and resources.

More importantly, in real-world applications, images are usually associated with a large number of labels [7]. For example, shown in Figure 1, we can see that each image can be labeled with several labels simultaneously, such as *beach*, *mountain*, *fall foliage*, *urban*, and *field*. It leads to multi-label image classification, which further increases the cost of constructing an effective training set. Thus, the task of utilizing abundant unlabeled images and limited resources to obtain an effective training set for training a high-performance multi-label classifier has become one of the most important issues in image classification.

To make full use of unlabeled images, two major machine learning approaches are proposed, namely, semi-supervised learning [3, 4, 12, 75, 89, 118, 122] and active learning [21, 62, 69, 85, 88]. Although both approaches address the above problem by training a classifier using a few of labeled images and a large number of unlabeled images, a fundamental difference between them is whether additional labeling of images by experts is required during training. Specifically, semi-supervised learning does not need manual intervention and directly utilizes unlabeled images to improve the performance of the classifier [122]. However, it is difficult to guarantee the automatic annotation accuracy of semi-supervised learning. Active learning, however, acquires the labels for some unlabeled images from domain experts, which can ensure accurate annotation. With limited resources, active learning iteratively selects the most informative examples to acquire their labels and trains a classifier from the updated training set, which is augmented with the newly selected examples. Active learning not only significantly reduces the labeling workload, but also constructs an effective training set for training a high-performance classifier. Therefore, active learning is a very useful and adoptable solution for multi-label image classification studies.

Due to the effectiveness of active learning, it has been widely used with a wide variety of approaches proposed in the past decades. Settles et al. [69] and Fu et al. [21] provided a general introduction and survey of recent developments in the active learning community. The recent developments in multimedia annotation, text classification, and remote sensing image classification can be found in References [62, 85, 88]. Wang et al. [88] mainly reviewed the integration of active learning with other related technologies, and Olsson et al. [62] and Tuia et al. [85] concentrated on the applications of active learning in natural language processing and remote sensing image classification, respectively. Existing surveys [21, 62, 69, 85, 88] mainly concentrated on general active learning or reviewed the applications of active learning in some specific fields. However, the discussion on multi-label active learning is limited and uncomprehensive without in-depth analysis. Moreover, the studies on multi-label active learning are developing rapidly, and there are many new advances of multi-label active learning over recent years. Thus, it is valuable and necessary to present a comprehensive and deep overview in this field for potential researchers. Also, future inspiring research directions are expected by the multi-label image classification community.

In this work, a more comprehensive survey focusing on multi-label active learning for image classification is presented. As described in the following sections, sampling and annotation are two essential components of an active learning system. Existing approaches for multi-label active learning will be categorized and analyzed from the two aspects. Furthermore, we will illustrate the future research directions in this field in terms of four potential aspects, i.e., example dimension, label dimension, annotation, and application extension.

The remainder of this work is organized as below. In Section 2, we briefly describe active learning and explain its general framework. In Section 3, we analyze the family of existing multi-label active learning algorithms and discuss the characteristics of representative algorithms. In Section 4, several future promises on multi-label active learning for image classification are given. Finally, we conclude this work in Section 5.

2 PROBLEM DEFINITION

According to various ways of querying, active learning algorithms can be categorized into three paradigms: membership query synthesis active learning [4, 45, 46, 48, 69], stream-based selective active learning [14, 16, 22, 47, 69, 109], and pool-based active learning [15, 50].

As we discussed, there exist a great amount of unlabeled images, and the number of images still keeps increasing at an exponential rate. Among these three paradigms, pool-based active learning is the most applicable one for practical scenarios, especially for image classification. Pool-based active learning has become the mainstream method, and many active learning algorithms for image classification were derived. Thus, this work concentrates on the analysis and discussion of pool-based active learning.

Figure 2 illustrates a general framework of a typical pool-based multi-label active learning system, which is an iterative cycle. *Sampling* and *Annotation* are two important components of a sample selection engine, as shown in the red-dotted boxes in Figure 2. Specifically, in each iteration, a multi-label classifier is trained by the current training set. Then, the informativeness of each example/example-label pair in the unlabeled data pool is assessed one by one using a designed informativeness measure. The most informative ones are selected, and a human expert is asked to label these selected examples/example-label pairs. After that, the training set is updated when new examples/example-label pairs are added. Sequentially, a new multi-label classifier is trained on the augmented training set. The whole cycle is iterative and stops when reaching a predefined stopping criterion, either the number of iterations or the performance of the classifier. To ensure that the description is intelligible and vivid, some definitions and symbols are presented as below.

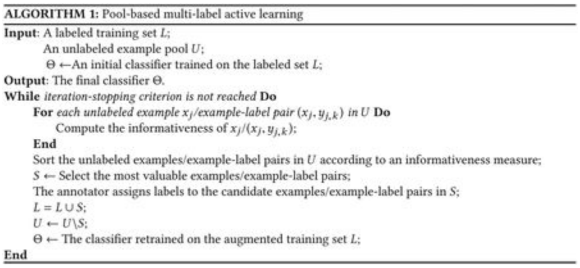
Given a multi-label active learning problem, there is an example set $X = \{x_1, \ldots, x_n\}$ consisting of n examples. Each example x_i belongs to a d-dimensional feature space and is associated with the possible label set Y_i, where Y_i is a subset of the total label set $A = \left\{l_1, l_2, \ldots, l_{n_m}\right\}$ and n_m is the number of labels in A. Besides, the definitions of a multi-label example, an example-label pair and informativeness are given as follows.

Definition 1. Multi-label example: A multi-label example x_i can be denoted as $\{x_i, Y_i\}$, where $x_i = \left\{v_1^i, v_2^i, \ldots, v_d^i\right\}$, and v_d^i denotes the dth feature value of x_i. $Y_i = \left\{y_{i, l_1}, y_{i, l_2}, \ldots, y_{i, l_{n_m}}\right\}$ denotes the label set of x_i.

Definition 2. Example-label pair: An example-label pair can be denoted as $\left(x_i, y_{i,\,l_s}\right)$, where $x_i = \left\{v_1^i, v_2^i, \ldots, v_d^i\right\}$ and $y_{i,\,l_s}$ denotes the *s*th label of x_i and $y_{i,\,l_s} \in \{0, 1\}$ is contained in the label set Y_i of example x_i. $y_{i,\,l_s} = 0$ indicates that the label $y_{i,\,l_s}$ does not belong to the example x_i. $y_{i,\,l_s} = 1$ indicates that the label $y_{i,\,l_s}$ belongs to the example x_i.

Definition 3. Informativeness: The measurement of useful information of an unlabeled example x_i or example-label pair $\left(x_i, y_{i,\,l_s}\right)$.

In a typical multi-label active learning scenario, there is an example set X including a small labeled training set $L = \{(x_i, Y_i) \mid 1 \le i \le n_l\}$ and a large unlabeled dataset $U = \{(x_j, Y_j) \mid 1 \le j \le n_u\}$. Here, n_l is the number of examples in L, and n_u is the number of examples in $U\,(n_l \ll n_u)$. Its ultimate goal is to learn a high-performance multi-label classifier $\Theta : X \rightarrow A$. The pseudo code is shown in Algorithm 1.



```
ALGORITHM 1: Pool-based multi-label active learning
Input: A labeled training set $L$;
       An unlabeled example pool $U$;
       $\Theta$ ←An initial classifier trained on the labeled set $L$;
Output: The final classifier $\Theta$.
While iteration-stopping criterion is not reached Do
    For each unlabeled example $x_j$/example-label pair $(x_j, y_{j,k})$ in $U$ Do
        Compute the informativeness of $x_j/(x_j, y_{j,k})$;
    End
    Sort the unlabeled examples/example-label pairs in $U$ according to an informativeness measure;
    $S$ ← Select the most valuable examples/example-label pairs;
    The annotator assigns labels to the candidate examples/example-label pairs in $S$;
    $L = L \cup S$;
    $U \leftarrow U \backslash S$;
    $\Theta$ ← The classifier retrained on the augmented training set $L$;
End
```

In next section, a summary of representative works on multi-label active learning is provided and their sample selection is discussed from two aspects, sampling and annotation. For ease of reference, Table 1 lists the major mathematical notations throughout this work along with their explanations. It first summarizes the basic notations defined above and also provides commonly used notations in the following section.

3 OVERVIEW OF TOPICS

As mentioned, effective solutions on constructing an effective training set for multi-label image classification are highly in demand in real-world applications. Many multi-label active learning algorithms have been proposed in recent years. In this section, a broad overview of representative sampling strategies of multi-label active learning is presented. Figure 3 shows the hierarchical structure of existing multi-label active learning algorithms. Two main areas of focus for these algorithms are sampling and annotation. Section 3.1 of this section gives a broad overview of representative sampling strategies and Section 3.2 focuses on annotation algorithms.

3.1 Sampling

To design a sampling strategy, two aspects should be considered, i.e., the selection of sampling granularity and the design of information measure. In this subsection, existing multi-label active learning algorithms for image classification will be analyzed and summarized from the two aspects.

3.1.1 Sampling Granularity.

—Sampling granularity indicates the selection unit of a sampling strategy of multi-label active learning. According to various selection units, we categorize these existing multi-label active learning algorithms into four types: example-based, example-label-based, mixed-mode-based and batch-mode-based methods.

A. Example-based Methods.

(a) Method Definition.: **Definition 4. Example-based:** Given an unlabeled dataset U and a known label space A, an example-based method selects n_s most informative examples $\{x^*\}_1^{n_s}$ from U and then for each example x^* acquires all the labels from domain experts. The informativeness of each unlabeled example can be calculated through an informativeness measure function $Info(\cdot)$. Accordingly, the sampling strategy of example-based methods can be defined as follows:

$$x^* = \underset{x_j \in U}{\arg\max}\ Info(x_j)\,. \tag{1}$$

For each unlabeled example, its informativeness might be assessed from several aspects, such as its uncertainty, representativeness, diversity, and other informativeness measures. The overall informativeness of an example is an integration of potential information measures. Since the influences of different information measures for model training are disparate, tradeoff parameters are used to balance the relative importances of these measures. Thus, the informativeness measure function $Info(\cdot)$ of an unlabeled example x_j can be denoted as follows:

$$\begin{aligned} & Info(x_j) = \sum \alpha_s I_s(x_j) \\ & s.t.\ \alpha_s \in [0, 1], \sum \alpha_s = 1, \end{aligned} \tag{2}$$

where $I_s\,(x_j)$ is one of the information measure methods and α_s is the sth tradeoff parameter.

In an example-based algorithm, all the labels of the selected examples are supposed to be annotated simultaneously. As an example, Figure 4 illustrates one iteration of an example-based multi-label active learning method. There are six images $\{x_1, \ldots, x_6\}$ and the label space A contains five different labels $\{$*Beach*, *Mountain*, *Fall Foliage*, *Urban, Field*$\}$. In the bottom two tables of Figure 4, "1" indicates that the label is assigned to this example, "0" indicates that the label is not assigned to this example, and "?" indicates that the label is unknown. Supposing that examples x_1, x_3, and x_6 are selected, all the labels of these three examples will be labeled according to the idea of the example-based sampling strategy.

Different sampling strategies of example-based multi-label active learning methods are proposed in References [10, 20, 34, 35, 38, 39, 52, 54, 60, 66, 68, 73, 78, 86, 102]. Boutell et al. [7] first proposed to solve a multi-label classification problem by decomposing it into a group of binary classification problems and a framework of multi-label scene classification was built in this work. It is known as problem transformation [83], which inspires many subsequent multi-label active learning studies, such as Li et al. [54], Li and Guo [52], and Singh et al. [73]. They constructed a set of standard support vector machines (SVMs

[27–29]) to conduct multi-label active learning, which trained a set of binary classifiers associated with corresponding labels. Li et al. [54] was the first to introduce active learning into multi-label image classification.

Clearly, acquiring all the labels of the selected examples is usually costly, especially when the number of labels that need to be annotated in an image is large. Besides, the informativeness of each unlabeled example is only assessed from the aspect of the example dimension. Inherent relationships buried in the label distribution are ignored, while there must be intrinsic relationships between labels, i.e., label correlation. Intuitively, label correlation is beneficial for enhancing the sampling strategy of multi-label active learning [115]. Example-label-based methods usually choose a subset of labels associated with a certain example during sampling by exploring label correlation. More information about label correlation will be described in Section 3.1.2B.

(b) Parameter Setting.: As mentioned in Equation (2), to optimally integrate different information measures together, tradeoff parameters are adopted to balance their relative importance to obtain the optimal integration scheme. Generally, there are two methods to determine suitable parameters.

One is to set fixed parameters based on experience and constraint conditions. Specifically, a set of different parameter groups are predefined first. This set is used to form a series of corresponding informativeness measure functions, which are used to train a corresponding set of classification models. Finally, a parameter group, which was used by the model that achieved the highest classification performance, is chosen as the final parameter setting. This method is used in many multi-label active learning studies. However, it is very rigid and has to be adapted for each new dataset, which is inconvenient and costly.

Another way is to develop an adaptive parameter tuning scheme to actively select a best integration parameter group for the dataset to achieve the best performance of the classifier [100]. For example, Li and Guo [52] developed a consistent but flexible parameter tuning strategy based on the expected loss to fit into different learning scenarios. More precisely, the sampling strategy for assessing the informativeness of each unlabeled example x_j is denoted as

$$Info\left(x_j\right) = I_1\left(x_j\right)^{\alpha} \cdot I_2\left(x_j\right)^{1-\alpha}, \tag{3}$$

where $\alpha \in [0, 1]$. Since it is hard to make continuous selection of α value, they selected α value from a predefined set of discretely sampled values, e.g., $V = [0, 0.1, \dots, 0.9, 1]$. For each α in the given set V, a series of corresponding informativeness measure functions is obtained. Using these functions, a batch of unlabeled examples (no more than $|V|$ examples) is selected from the unlabeled data pool and collected into a candidate set S for each, respectively. For each example x' in S, they used a multi-label SVM classifier $\Theta^0 = \left[f_1^0, \dots, f_{n_m}^0\right]$ trained on the current training set L to predict its label vector Y'. Then, a new multi-label SVM classifier $\Theta = \left[f_1, \dots, f_{n_m}\right]$ is trained on the augmented training set $L \cup (x', Y')$.

The approximate generalization error of the new classifier induced by the candidate example x' is defined as follows:

$$\varepsilon(x') = \sum_{j=1}^{n_u} \max_{p \in Y_j^+} \left[1 - f_p(x_j)\right]_+ + \max_{q \in Y_j^-} \left[1 + f_q(x_j)\right]_+, \tag{4}$$

where Y_j^+ denotes the predicted positive label set of the unlabeled example x_j by the classifier Θ and Y_j^- denotes the predicted negative label set correspondingly. Finally, the parameter group selection is defined as follows:

$$\alpha^* = \underset{\alpha_k \in V}{\arg\min} \sum_{x' \in S_k} \varepsilon(x'), \tag{5}$$

where S_k is the candidate set corresponding to α_k.

B. Example-Label-based Methods.

(a) Method Definition.: **Definition 5. Example-label-based:** Given an unlabeled example set U and a known label space A, an example-label-based method selects n_s most informative example-label pairs $\{(x, y)^*\}_1^{n_s}$ from U and then for each example-label pair $\{(x, y)^*\}$ acquires corresponding label from domain experts. Here, n_s is the number of the selected example-label pairs, $(x, y)^*$ is the selected example-label pair, and y is the label contained in the label set of the example x. The informativeness of each example-label pair is calculated by the informativeness measure function $Info(\cdot)$. Accordingly, the sampling strategy of an example-label-based method is written as

$$(x, y)^* = \underset{x_j \in U, y_{j,k} \in UL(x_j)}{\arg\max} Info(x_j, y_{j,k}). \tag{6}$$

Similarly, for each unlabeled example-label pair, its informativeness is an integration of various information measures. Thus, the informativeness measure function $Info(\cdot)$ of an unlabeled example-label pair $(x_j, y_{j,k})$ is denoted as follows:

$$\begin{aligned} &Info(x_j, y_{j,k}) = \sum \alpha_s I_s(x_j, y_{j,k}) \\ &s.t.\ \alpha_s \in [0, 1], \sum \alpha_s = 1. \end{aligned} \tag{7}$$

where $I_s(x_j, y_{j,k})$ is one of the information measures and α_s is the sth tradeoff parameter.

Note that only part of labels rather than all the labels of candidate examples need to be labeled in an example-label-based method. For example, there are six images shown in Figure 5, and the images are the same with those shown at the top of Figure 4. Here, we assume that the number of labels needed to be annotated is the same as that required in Figure 4. Figure 5 illustrates the labels obtained from domain experts based on an example-label-based method. As we can see from Figure 5, for the example x_1, only the labels

{*Beach*, *Fall Foliage*, *Urban*, *Field*} are labeled, which corresponds to the example-label pairs $(x_1, y_{1,1})$, $(x_1, y_{1,3})$, $(x_1, y_{1,4})$, and $(x_1, y_{1,5})$.

Comparing to example-based methods, example-label-based methods can dramatically reduce the annotation cost, since they only need to acquire part of labels of an unlabeled example. Moreover, it is beneficial to explore underlying information buried in the labels. That is why example-label-based methods become the mainstream of multi-label active learning algorithms, and many relevant works [24, 30, 33, 63, 64, 91, 93–98, 105, 106, 111, 112, 114–116] have been proposed.

For instance, Qi et al. [63, 64] took into account not only the example space but also the label space and proposed a multi-label active learning method based on two dimensions. Wu et al. [94] also intended to acquire the labels for the most uncertain example-label pairs, which is the first work to propose the specific concept of example-label-based in multi-label active learning. To speed up the whole active learning process, Zhang et al. [111, 112] further proposed a batch mode for the selection of informative example-label pairs during each iteration. Guo et al. [30, 91] combined low-rank mapping with multi-label active learning to mine label correlation for improving the performance of the classification model.

However, to select high-informativeness example-label pairs, an example-label-based method needs to assess the overall informativeness of all the example-label pairs in the unlabeled set U. Its total computational cost is strongly affected by the number of labels, especially on a dataset with a large number of labels. Therefore, to develop an efficient example-label-based method also needs to take computation optimization into account.

(b) Parameter Setting.: Similarly to the example-based methods, the example-label-based methods also utilizes tradeoff parameters to integrate multiple information measures. The two parameter setting ways of the example-based methods can also be extended for the example-label-based methods. The fixed-weights setting for parameters is a commonly used approach in the example-label-based studies while the automatic parameter tuning scheme is a more effective way. For instance, Wu et al. [91] developed an adaptive integration framework based on the expected loss to automatically balance relative importance degrees among three information measures.

C. Mixed-Mode-based Methods.: **Definition 6. Mixed-mode-based:** Given an unlabeled example set U and a known label space A, a mixed-mode-based method selects n_s most informative examples $\{x^*\}_1^{n_s}$ from U first, and then for each candidate example x^* selects an optimal label subset Y_{sub}^* to acquire corresponding labels from domain experts. The final selection can be denoted as $\{x^*, Y_{sub}^*\}_1^{n_s}$. Here, n_s is the number of the selected example-label pairs, x^* is the selected example, and Y_{sub}^* is the selected label set contained in the label set Y^* of x^*. The informativeness of each example is calculated by the information measure function $Info_1(\cdot)$ and the informativeness of each label of the candidate example is calculated by the information measure function $Info_2(\cdot)$. Accordingly, the sampling strategy of a mixed-mode-based method is written as follows:

$$x^{*} = \underset{x_j \in U}{\arg\max}\, Info_1(x_j), \tag{8}$$

$$Y^{*}_{sub} = \underset{y_{*,k} \in UL(x^{*})}{\arg\max}\, Info_2(y_{*,k}). \tag{9}$$

Generally, the most informative examples are first selected based on an example-based sampling strategy. Then, for each candidate example, part of labels will be selected based on a label-based selection strategy to construct an optimal label subset. Finally, the selected examples and labels are combined to augment the training set after annotated by human experts. As we can see, a mixed-mode-based method is a balanced decision. In fact, it is a special case of the example-label-based method.

For example, Jiao et al. [40] proposed a typical mixed-mode-based multi-label active learning method, max-margin uncertainty sampling with label-set push (MUSLAP). First, it selects a batch of the most uncertain examples based on a max-margin-based uncertainty sampling strategy. Then, it determines an optimal label set of each candidate example by estimating its label concept composition as the label concept composition reflects the probability distribution of the example with different labels.

Besides, Huang et al. [36] also introduced this mixed mode to design their sampling strategy. In their study, they also utilized uncertainty to select candidate examples, where the uncertainty of each example depends on the number of performed queries. The fewer queries performed, the more uncertain the example is. That is, the fewest queried examples will be selected at first. Then, for each candidate example, two certain labels (as a label pair) whose queried information induces a large change to the classification model will be selected. Here, the label pair is supposed to have two properties. First, their relevance ordering should be less confident by the current model. Second, the prediction of the two labels should not be too close.

Comparing to example-based and example-label-based methods, the mixed-mode-based methods look like the latter one, since they also consider both the example dimension and the label dimension. Meanwhile, they can also reduce the annotation cost, since only a small label set of each candidate example needs to be annotated.

D. Batch-Mode-based Methods.: Most state-of-the-art multi-label active learning methods select one unlabeled example at a time. However, to reduce the waiting time of annotations, selecting batches of unlabeled examples is preferred. A simple and commonly used way of picking a batch of unlabeled examples is to select the best unlabeled examples by following a greedy approach. However, the main drawback of this solution lies in the redundancy of information. To overcome this drawback, the *batch-mode active learning* approach is used: it allows selecting a batch of unlabeled examples, which are informative and have minimal redundancy.

To date, the study of batch-mode active learning on multi-label data is very scarce [10, 24, 38, 39, 68, 91, 111, 112]. Chakraborty et al. [10] proposed a batch-mode multi-label active learning strategy, which queries all the labels of the selected unlabeled examples from human experts. The informativeness of an unlabled example was computed as the average entropy over individual labels. The diversity between unlabeled examples was computed using a matrix that contains the redundancy values between each pair of unlabeled examples. Jiao et al. [38, 39] proposed a multi-criterion query-based batch-mode active learning technique. Specifically, after clustering a pre-selected uncertain set using a kernel *k*-means clustering algorithm, a Gaussian process was used to select the most informative sample in each cluster. Zhang et al. [111, 112] proposed a high-order label correlation driven active learning strategy. After selecting unlabeled examples based on an example-label pairs scheme, the batch-mode multi-label active learning problem was formulated as an NP-hard integer programming problem with constraints. Gao et al. [24] proposed a batch-mode strategy named multi-label active learning based on distribution matching. Specifically, the representativeness and diversity criteria were used to select a batch of unlabeled examples according to labeled and unlabeled sets with regards to their marginal probability distributions in feature and label spaces. Reyes et al. [68] formulated the batch-mode multi-label active learning as a multi-objective problem and resolved this problem using an evolutionary algorithm. The proposed strategy intends to optimize three criteria at the same time (informativeness, representativeness and diversity) to select the best batch of unlabelled examples. It also has a lower computational cost than previous batch-mode active learning methods do, allowing its application to large multi-label datasets.

The representativeness and diversity are good criteria for improving the generalization of base classifiers when used for batch-mode active learning methods. Representativeness-based criteria measure whether an unlabeled example is representative of the underlying distribution by analyzing the density around examples in the feature space. Diversity-based criteria measure the information overlap among a set of examples in the label space. The representativeness and diversity computation is described in detail in Sections 3.1.2C and 3.1.2D, respectively.

3.1.2 Informativeness Measure.—As discussed, a sampling strategy is the key of a multi-label active learning algorithm, which aims at labeling the examples/example-label pairs with high informativeness. An information measure for assessing the informativeness of each example is the core of a sampling strategy. Considering that there is much potential information buried in the example distribution and label distribution, it is helpful to mine these kinds of information as basic information measures.

In this subsection, we illustrate different basic information measures used in existing multi-label active learning algorithms. Here, we categorize these basic informativeness measures into six groups, as shown in Figure 3:

1. Uncertainty metric;
2. Label correlation metric;
3. Representativeness metric;

4. Diversity metric;
5. Noise content metric;
6. Expected error reduction metric.

A. Uncertainty Metric.: **Definition 7. Uncertainty metric:** Given an unlabeled set U and a known label space A, an uncertainty metric is a function Unc used to measure the uncertainty of a classifier to predict an unlabeled example/example-label pair. The uncertainty metric in an example-based active learning method is calculated based on the unlabeled set U, while it is calculated from both examples and labels $U \times A$ in an example-label-based active learning method:

$$Unc : \begin{cases} U \mapsto \mathscr{R} & example - based \\ U \times A \mapsto \mathscr{R} & example - label - based, \end{cases} \tag{10}$$

where $\mathscr{R}$ defines a real number space. Uncertainty sampling aims to select an example/example-label pair where the current classifier is the most uncertain in its prediction, which is a simple and widespread used strategy. Many multi-label active learning algorithms adopted the uncertainty metric in their sampling strategies.

As it is easy to obtain prediction probabilities or confidence scores from classifiers, a simple uncertainty metric is easily proposed with an underlying intuition that the lower the prediction confidence, the higher informativeness the corresponding examples/example-label pairs have to improve the current classifier [36]. For example, Singh et al. [73] trained several SVM classifiers and calculated the SVM margin values for all labels of an example simultaneously. By converting the distances from the margins to probability scores, the average probabilities across all SVMs can identify the uncertainty of the example.

Combining the classification margin value with a simplest rank aggregation method, Reyes et al. [66] proposed a score function to measure the uncertainty of each unlabeled example. Specifically, they first calculated the margin value for utilizing the predictions from the current classifier to represent whether an example belongs to a certain label. Note that a larger margin value means that the classifier has a smaller prediction error. Accordingly, the margin value mar_j of the label $y_{j,k}$ in the example x_j is written as follows:

$$mar_j(y_{j,k}) = |p(y_{j,k} = 1 \mid x_j) - p(y_{j,k} = 0 \mid x_j)|. \tag{11}$$

Using Equation (11), they obtained a vector of margin values $M(x_j) = [mar_j(y_{j,1}), \ldots, mar_j(y_{j,n_m})]$ for each unlabeled example x_j. Then, after collecting and listing the margin values of all the unlabeled examples on the label $y_{j,k}$, they computed a ranking $\tau_{y_{j,k}}$ to order the unlabeled examples as follows:

$$\tau_{y_{j,k}} = (x_1, x_2, \ldots, x_{n_u}) \mid mar_1(y_{j,k}) < \cdots < mar_{n_u}(y_{j,k}). \tag{12}$$

After that, they adopted the Borda’s method, which is a positional method to assign a score to an element in correspondence to the position where this element appears in each ranking, to compute the score of each unlabeled example x_j as

$$Unc\left(x_j\right) = \frac{\sum_{y_{j,k} \in Y_j}\left(n_u - \tau_{y_{j,k}}\left(x_j\right)\right)}{n_m\left(n_u - 1\right)}. \tag{13}$$

Several works followed the same idea above, such as Huang et al. [34, 35].

Certainly, there are some other ways to represent the uncertainty based on the classifiers’ prediction probabilities. The simplest and most widely used method used in References [20, 30, 93–96, 105, 106] can be denoted as follows:

$$Unc\left(x_j, y_{j,k}\right) = \left|\frac{1}{2} - p\left(y_{j,k} = 1 \mid x_j\right)\right|, \tag{14}$$

where $p(y_{j,k} = 1|x_j)$ defines the prediction probability that the label $y_{j,k}$ is associated with the example x_j. Tur et al. [86] made a few alterations and predefined a threshold as a lower prediction probability limit to query the example-label pair whose prediction probability conforms to the criteria.

Entropy is another commonly used uncertainty measure method [91, 97, 98, 111, 112, 115, 116]. Within example-label-based methods, it can be represented as follows:

$$Unc\left(x_j, y_{j,k}\right) = -\sum_{m \in \{0,1\}} p\left(y_{j,k} = m \mid x_j\right) \log p\left(y_{j,k} = m \mid x_j\right). \tag{15}$$

Moreover, to combine entropy with mutual information, Qi et al. [63, 64] achieved a new reliable uncertainty metric, which is able to explore the uncertainty on both the example and label dimensions. Specifically, this two-dimensional active learning approach is denoted as follows:

$$\begin{aligned} Unc\left(x_j, y_{j,k}\right) &= \sum_{s=1}^{n_m} MI\left(y_{j,s}; y_{j,k} \mid y_{j,LD\left(x_j\right)}, x_j\right) \\ &= H\left(y_{j,k} \mid y_{j,LD\left(x_j\right)}, x_j\right) + \sum_{s=1, s \neq k}^{n_m} MI\left(y_{j,s}; y_{j,k} \mid y_{j,LD\left(x_j\right)}, x_j\right), \end{aligned} \tag{16}$$

where the former one $H(\cdot)$ is the uncertainty measured from the selected example-label pair $(x_j, y_{j,k})$ itself, and the latter one $MI(\cdot)$ is the statistical redundancy between the selected label $y_{j,k}$ and the rest $y_{j,LD\left(x_j\right)}$. The label uncertainty is decreased when the inference of other labels can obtain some information by maximizing the mutual information term. Furthermore, Zhang et al. [114] considered the uncertainty over different views when selecting example-label pairs and extended this algorithm under a multi-view scenario.

To take into consideration the uncertainty on both the example and label dimensions, Li and Guo [52] proposed two metrics to assess the uncertainty on the two dimensions, respectively.

First, a max-margin prediction uncertainty selection strategy is designed from the example dimension, which is also used in Reference [40]. As we know, based on SVM [27–29], it is clear that the separation margin of the most uncertain example is the smallest as multi-label classification is actually about the overall separation between the groups of positive labels and negative labels. Accordingly, a max-margin uncertainty measure is modeled using a global separation margin, which is represented by an inverse separation margin as follows:

$$Unc\left(x_j\right) = \frac{1}{\min_{k \in \hat{Y}_j^+}\left|f_k\left(x_j\right)\right| + \min_{s \in \hat{Y}_j^-}\left|f_s\left(x_j\right)\right|}, \tag{17}$$

where the denominator is the separation margin over example x_j, $\hat{Y}_j^+$ and $\hat{Y}_j^-$ denote the predicted positive and negative labels of an unlabeled example x_j of a corresponding classifier, respectively, and f_k is the SVM classifier for the kth label trained from the segmented labeled set $L \cup (x_j, Y_j)$.

Then, an inconsistency strategy based on label cardinality [81] is designed from the label dimension. Since all the examples are extracted from the same natural distribution, the statistical properties should be shared by the input features and the labels simultaneously. Accordingly, Euclidean distance is chosen as the label inconsistency, which calculates the distance between the label cardinality of the current labeled data and the number of predicted positive labels,

$$Unc\left(x_j\right) = \left\| \sum_{k=1}^{n_m} I_{\left[y_{j,k} > 0\right]} - \frac{1}{n_l} \sum_{s=1}^{n_l} \sum_{k=1}^{n_m} I_{\left[y_{s,k} > 0\right]} \right\|_2, \tag{18}$$

where $I_{[\cdot]}$ expresses an indicator function. When the given condition is met, it is 1; otherwise, 0.

Besides, the conflicts between the known label predictions and the obtained label correlation can also be regarded as an uncertainty metric, which is known as cross label uncertainty [66, 111, 112, 115]. Specifically, it is achieved by summing all the Kullback-Leibler divergences between each label $y_{j,k}$ and its correlated labels in example x_j, as defined:

$$Unc\left(x_j, y_{j,k}\right) = \frac{1}{\left|c\left(y_{j,k}\right)\right|} \sum_{y_{j,s} \in c\left(y_{j,k}\right)} D_{KL}\left(p_k \parallel p_s\right). \tag{19}$$

The Kullback-Leibler divergences D_{KL} is calculated as follows:

$$D_{KL}\left(p_k \parallel p_s\right) = -\sum_{y_{j,k} \in \{0,1\}} p\left(y_{j,k} \mid x_j, f_k^t\right) \log \frac{p\left(y_{j,k} \mid x_j, w_k^t\right)}{p\left(y_{j,s} \mid x_j, w_s^t\right)}, \tag{20}$$

where t denotes the tth iteration, w_k^t is a parameter vector of classifier for the label $y_{j,k}$ learned from the training set in the tth iteration, $p\left(y_{j,k} \mid x_j, w_k^t\right)$ is the prediction probability of $y_{j,k}$ in x_j under w_k^t, and f_k^t is the classifier for the label $y_{j,k}$ trained in the tth iteration.

B. Label Correlation Metric.: Definition 8. Label correlation metric: Given an unlabeled set U and a label space A, a label correlation metric is a function Cor used to assess the relevance between any two labels as follows:

$$Cor: U \times A \mapsto \mathscr{R} \qquad \text{example-label-based.} \tag{21}$$

In practical applications, each image can be annotated by several specific labels. There exist intrinsic relationships among labels, such as co-occurrence, mutual exclusion, and independence. These correlations are capable of providing valuable information for the classifier training. For example, if a label $y_{j,k}$ belongs to an example x_j and the relationship of two labels $y_{j,k}$ and $y_{j,s}$ is co-occurrence, then it is more likely that the label $y_{j,s}$ also belongs to the example x_j. Thus, many studies adopted label correlation as an information measure metric.

Depending on its characteristics, label correlation can be distinguished into two different types, unconditional label dependence and conditional label dependence.

(a) Unconditional Label Dependence.: **Definition 9. Unconditional label dependence:** Unconditional label dependence is a global static label dependence existing in the label space, which is independent to concrete observations.

At present, unconditional label dependence is frequently used in multi-label active learning studies. For example, to characterize the label relationship, Jiao et al. [38] constructed a label hierarchy tree to describe the label correlation in a tree structure. The label hierarchy tree is built recursively in a top-down approach by depth-first traversal on the training set. The balanced k-means cluster algorithm [82] is used to segment the label set into several irrelevant subsets, which is corresponding to several child nodes. This process is repeated until each child node only contains one single label.

To take advantage of label correlation to enhance the training process, Zhang et al. [115] developed the measure for assessing the informativeness of an example-label pair by considering the information gain of its corresponding label on the related label set of this label. Specifically, based on the information gain, the label correlation of an example-label pair $(x_j, y_{j,s})$ is calculated as follows:

$$Cor(x_j, y_{j,k}) = \sum_{y_{j,s} \in Prop_C(y_{j,k}),\, y_{j,s} \in UL(x_j)} R(y_{j,s}, x_j), \tag{22}$$

where $R(\cdot)$ is the information gain of the label $y_{j,k}$ on one of its related labels $y_{j,s}$. $Prop_C(y_{j,k})$ is a set of propagated outcomes to represent the labeling outcome that they can be inferred from the assignment $y_{j,s}$ based on a set of constraints C, denoted as follows:

$$Prop_C(y_{j,k}) = \left\{ y_{j,s} \mid y_{j,k} \overset{C}{\rightarrow} y_{j,s} \right\}. \tag{23}$$

Therefore, the inference of outcomes is based on a series of rules provided by constraints. For example, the inheritance constraint “$y_{j,k}$ is a derived class of $y_{j,s}$” provides two rules “$y_{j,k} = 0 \rightarrow y_{j,s} = 0$,” and “$y_{j,k} = 1 \rightarrow y_{j,s} = 1$.”

Based on the same idea, Nasierding and Kouzani [60] assessed label correlation by measuring the rank of label predictions, i.e., ranking the most relevant label to obtain the highest score and ranking the most irrelevant one to obtain the lowest score. Therefore, average precision evaluating the average fraction of labels ranked above a particular label $y_{j,k}$ is defined as follows:

$$Cor(x_j, y_{j,k}) = \frac{1}{n_u}\sum_{j=1}^{n_u}\frac{1}{\left|LD(x_j)\right|}\sum_{y_{j,k} \in LD(x_j)}\frac{\left|y_{j,s} \in LD(x_j): r_j(y_{j,s}) \le r_j(y_{j,k})\right|}{r_j(y_{j,k})}, \tag{24}$$

where $r_j(y_{j,k})$ is the label ranking of the label $y_{j,k}$. Similarly, Huang et al. [36] also proposed to minimize an approximated rank loss to rank relevant labels before all irrelevant ones. They trained a dummy label to separate relevant and irrelevant labels from the ranked label list.

Furthermore, unconditional label dependence can be explored through statistical analysis as well. Some relevant statistical approaches are introduced. For instance, association rule mining [1] is a useful method to discover frequently co-occurred label pairs, which aims to obtain a series of underlying association rules to acquire the correlated label set of a certain label [111, 112]. Specifically, considering a known example with its labeled label set as a transaction, two measurements, *support* and *confidence*, are required to conduct association rule mining:

$$\text{support }(l_k \Rightarrow l_s) = support(l_k \cup l_s), \tag{25}$$

$$confidence(l_k \Rightarrow l_s) = \frac{support(l_k \cup l_s)}{support(l_k)}, \tag{26}$$

where the *support* of l_k is equal to the percentage of the transactions that includes l_k,

$$support(l_k) = \frac{\left|\{t_i \mid l_k \subseteq t_i\}\right|}{T}, \tag{27}$$

where t_i is the ith transaction and T is the number of all the transactions. Combing these two measurements, the label correlation is obtained.

Besides, two methods, cosine similarity and chi-square statistics, are also popular in label correlation metric designing. For example, Ye et al. [105] proposed a cosine similarity-based multi-label active learning (CosMAL) method, and the correlation between the label pair (l_k, l_s) is denoted as follows:

$$c_{ks} = \cos\left(\overrightarrow{l}_k, \overrightarrow{l}_s\right) = \frac{\left\langle \overrightarrow{l}_k, \overrightarrow{l}_s \right\rangle}{\left\| \overrightarrow{l}_k \right\| \left\| \overrightarrow{l}_s \right\|}. \tag{28}$$

However, though this traditional method is useful, it only considers the positive label correlation while ignoring the negative one. The negative label correlation is more common in real-world applications. Incomplete label correlation exploration might have a negative impact on training a classification model. Thus, cosine similarity-based methods are more suitable to the dataset, whose number of positive-relation label pairs is large. To address this issue, Ye et al. [95, 96, 106] further put forward a CSMAL method for a full correlation estimation using the chi-square statistics [44]. Note that the correlation measured by the chi-square statistics is symmetrical. That is, $c_{ks} = c_{sk}$; $c_{ks} > 0$ indicates a positive correlation between the label pair (l_k, l_s), while $c_{ks} < 0$ indicates that their relationship is negative.

After obtaining the correlation between each possible label pair, the label correlation of each example-label pair can be calculated as follows:

$$Cor\left(x_j, y_{j,k}\right) = \frac{1}{\left|UL\left(x_j\right)\right|} \sum_{y_{j,s} \in UL\left(x_j\right)} \left|c_{ks}\right|. \tag{29}$$

The approaches mentioned above are purely based on the observed label distribution, which is actually extracted from an incomplete training set. Since only partial labels of the selected examples will be annotated in an example-label-based algorithm, the training set is incomplete. It is clear that it cannot provide sufficient and reliable information for a full and correct label correlation mining. To solve this problem, Wu et al. [91] proposed a solution to automatically assign values to the missing labels in an incomplete training set to have a complete training set, so that they could obtain more related information from this more complete label distribution. Then, they further adopted chi-square statistics to obtain the label correlations based on the completed training set.

(b) Conditional Label Dependence.: **Definition 10. Conditional label dependence:** Conditional label dependence is a special label dependence that only exists in a specific set of examples.

Zhao et al. [116] and Wu et al. [97] are the first two articles concentrating on utilizing conditional dependence in multi-label active learning. Having observed that adding relevant label pairs to training set benefits to the performance improvement of a classifier, conditional label dependence between each label pair can be evaluated based on the gain for binary classification after taking advantage of this conditional dependence. More precisely, it is necessary to learn four binary classifiers for each label pair, respectively. Two of them are only trained from the original example space with each label from the label pair as the target, respectively. For simplicity, we simply call these two classifiers as original classifiers. After augmenting the original example space with each label from the label pair, two more classifiers are trained with the other label as the target. These two classifiers are called augmented classifiers. The accuracy of each classifier is obtained using *k*-fold cross-validation. Comparing to the original classifiers, if the two augmented classifiers have significantly higher accuracies respectively, then this indicates that there is conditional dependence among these label pair. To determine the statistical significance of the average accuracy difference among each classifier pair of classifiers, a paired *t*-test is adopted, which is denoted as follows:

$$t = \frac{\bar{a}_1 - \bar{a}_2}{\sqrt{\frac{1}{2}\left(s_{a_1}^2 + s_{a_2}^2\right)} \times \sqrt{\frac{2}{n_u}}}, \tag{30}$$

where $\bar{a}_i$ expresses the average value of the classifiers' accuracies and $s_{a_1}^2$ expresses the unbiased estimation of the two accuracies' variances. While the correlated label sets are obtained, they introduced information gain using *KL* divergence as the label correlation metric, denoted as follows:

$$Cor\left(x_j, y_{j,k}\right) = \sum_{y_{j,s} \in c\left(y_{j,k}\right)} \sum_{m \in \{0,1\}} p_k\left(y_{j,k} = m \mid x_j\right) \ln \frac{p_k\left(y_{j,k} = m \mid x_j\right)}{p_s\left(y_{j,s} = m \mid x_j\right)}, \tag{31}$$

where $p_k\left(y_{j,k} | x_j\right)$ is the posterior probability, denoting the probability of label $y_{j,k}$ belonging to example x_j.

Furthermore, Wu et al. [98] proposed the concept of asymmetrical conditional dependence depending on the difference of label dependency degrees. To assess the label dependency degree, they intended to normalize all the dependence degrees to a region [0, 1] via a row-based normalization. Given a conditional dependence matrix *R*, each element r_{ks} presents the dependence degree that label l_k depends on label l_s. Each row of *R* represents the dependence degrees of all labels that depends on the label l_s. Specifically, they normalized the dependence degree in *R* to obtain a normalized dependence matrix *W*, through the following normalization formula:

$$\begin{cases} w_{ks} = \dfrac{r_{ks}}{\sum_{s=1}^{k-1} r_{ks} + \sum_{s=k+1}^{n_m} r_{ks}} & (k \neq s) \\ w_{ks} = 1 & (k = s) \end{cases}, \tag{32}$$

where w_{ks} is the normalized dependence degree of the label l_k depending on the label l_s. The label dependence degree for the label with itself is set as $w_{ks} = 1$. If $w_{ks} = 0$, then it means that the label l_s is completely independent to the label l_k.

C. Representativeness Metric.: Definition 11. Representativeness metric: Given an unlabeled set *U* and a known label space *A*, a representativeness metric is a function *Rep* used to estimate the representativeness of an unlabeled example in the whole unlabeled dataset *U* or an unlabeled example-label pair in its corresponding example set:

$$Rep : \begin{cases} U \mapsto \mathscr{R} & \text{example-based} \\ U \times A \mapsto \mathscr{R} & \text{example-label-based} \end{cases}. \tag{33}$$

Representativeness of an unlabeled example reflects what degree this example could represent others. Since there are some outliers in the dataset, they will mislead the classification model training. Selecting the most representative example from the unlabeled dataset can reduce the probability of adding the outliers to the training set and help a classifier to learn on an actual example distribution.

For instance, a representativeness metric can be developed based on the approximately correct decision boundary [34, 35]. More specifically, the mutual information between a candidate example x_j and the remainder unlabeled examples in U can be used to assess the representativeness of each unlabeled example x_j as well [39]. It is developed based on the Gaussian Process framework [42] and is defined as follows:

$$Rep(x_j) = I(x_j, U - \{x_j\}) = \frac{1}{2}\ln\left(\frac{\sum_{jj}}{\sum_{j \mid U_j}}\right), \tag{34}$$

where $U - \{x_j\}$ is the current unlabeled example set expect the example x_j. The covariance matrix is generated by a Gaussian Kernel function $K(x_j, x_s) = \exp\left(-\frac{\|x_j - x_s\|^2}{2\lambda^2}\right)$, which is symmetric positive definite, as follows:

$$\sum_{U_jU_j} = \begin{pmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_{n_u}) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_{n_u}) \\ \vdots & \vdots & \vdots & \vdots \\ K(x_{n_u}, x_1) & K(x_{n_u}, x_2) & \cdots & K(x_u, x_{n_u}) \end{pmatrix}, \tag{35}$$

where $\sum_{U_jU_j}$ is a kernel matrix defined over all the unlabeled examples indexed by U_j, and U_j is assumed as $U_j = \{1, 2, \ldots, n_u\}$.

Generally, we can see that the representativeness metric can avoid the drawbacks associated with the uncertainty metric and tend to gradually improve the classification model.

D. Diversity Metric.: Definition 12. Diversity metric: Given an unlabeled set U and a known label space A, a diversity metric is a function Div used to measure the diversity of an unlabeled example corresponding to the labeled example set or an unlabeled example-label pair corresponding to the labeled example-label pair set:

$$Div : \begin{cases} U \mapsto \mathscr{R} & \text{example-based} \\ U \times A \mapsto \mathscr{R} & \text{example-label-based} \end{cases}. \tag{36}$$

In other words, it is supposed to reduce the redundancy inside the labeled training set and maximize the generalization ability of the training set. To obtain the diversity, redundancy between examples/example-label pairs should be measured first.

Definition 13. Redundancy: Redundancy is a function *Red* used to measure the similarity between an example pair (x_j, x_s), which is called pairwise redundancy, defined as below.

$$Red(x_j, x_s) \Leftrightarrow Similarity(x_j, x_s) . \tag{37}$$

After uncovering pairwise redundancy between each example pair, we can estimate the diversity of an unlabeled example/example-label pair then.

To measure the redundancy, Joshi et al. [41] introduced a cosine similarity method. Specifically, based on a kernel K, the redundancy among an example pair (x_j, x_i) can be defined as follows:

$$Red(x_j, x_i) = \cos(\langle x_j, x_i \rangle) = \frac{|K(x_j, x_i)|}{\sqrt{K(x_j, x_j)K(x_i, x_i)}} . \tag{38}$$

Accordingly, the diversity of an example x_j is estimated as follows:

$$Div(x_j) = 1 - \max_{x_i \in L} Red(x_j, x_i), \tag{39}$$

where L is the current training set constructed by all the labeled examples.

There is something different in Reference [91], which regarded the diversity of each unlabeled example-label pair (x_j, $y_{j,k}$) as the average value of the redundancy between its example and the examples of all the labeled example-label pairs in the current training set, which is denoted as follows:

$$Div(x_j, y_{j,k}) = 1 - \frac{1}{n_l} \sum_{x_i \in L} Red(x_j, x_i), \tag{40}$$

where n_l is the number of the examples in the training set L.

In a multi-label active learning process, a classifier has to be trained iteratively, which is time-consuming. To speed up the process of active learning, batch-mode is introduced, which allows incremental addition of a batch of labeled examples/example-label pairs to the training set. The entire batch will be labeled by domain experts simultaneously. It is hard to directly optimize the selection of examples/example-label pairs under batch-mode as it is usually accompanied by a greedy process. However, it is more convenient to ensure that the selected unlabeled examples/example-label pairs are different from each other in the batch (a diverse batch). That is, the selection of each example/example-label pair depends on the sum of its diversity to others that are already in the batch.

Jiao et al. [38, 39] noticed that the batch of selected examples should be kept diverse and applied the kernel k-means clustering algorithm to select the most diverse examples. More precisely, during the clustering process of an example set, the most similar examples will be assembled in the same cluster. Then, the less representative example in each cluster is deemed to be the most diverse. Accordingly, the diversity sampling is defined as follows:

$$x^* = \underset{x_j \in C_i}{\arg\min}\, Rep(x_j), \tag{41}$$

where $Rep(x_j)$ is illustrated in Equation (33) and C_i is the ith cluster after clustering all the candidate examples.

E. Noise Content Metric.: Definition 14. Noise Content Metric: Given an unlabeled set U and a known label space A, a noise content metric is a function Noi used to measure the noise content of an unlabeled example/an unlabeled example-label pair to quantize the influence of noise on their original data distribution:

$$Noi : \begin{cases} U \mapsto \mathscr{R} & \text{example-based} \\ U \times A \mapsto \mathscr{R} & \text{example-label-based} \end{cases}. \tag{42}$$

Sample data are captured from real-world applications, and it is always influenced by noise. So, the classification model is usually learned from noisy data. Noisy data will mislead the estimation of the data distribution and subsequently deteriorate the performance of the model built from itself. Thus, some researchers endeavored to explore how to reduce the interference of the noise.

(a) Denoising.: Without external influence, the subspaces of a dataset are low-rank, independent and noiseless. All of these will be destroyed when there exists noise, since the noise distribution is usually scattered randomly and does not follow original subspace structures. Considering this phenomenon, Liu et al. [57] proposed a low-rank representation (LRR) model to identify the original subspace structures from corrupted data by segmenting examples into their respective subspaces.

As the LRR algorithm is tractable with a good robustness to noise, Wu et al. [91] adopted this low-rank representation method into multi-label active learning to reduce the noise influence. They separated out the noise data and extracted a low-rank representation to represent the original data distribution instead of the noisy data for model training. More precisely, since the presence of noise will result in a high-rank feature space, they imposed a low-rank constraint to reduce the rank of the noisy data. When reaching the lowest rank, a low-rank representation will be obtained, which contains the original distribution of the noisy data. It can be written as follows:

$$\begin{aligned} &\min_{Z, E} rank(Z) + \lambda \parallel E \parallel_0 \\ &s.t.\ U = UZ + E, \end{aligned} \tag{43}$$

where Z is a low-rank transition matrix that is used to acquire a low-rank representation of the noisy data U, $rank(Z)$ denotes the rank of matrix Z, λ is a tradeoff parameter, E is the noise data that is represented by a sparse matrix, and $\|\cdot\|_0$ is the 0-norm. To preserve the data structure and approximately expose the underlying subspace of U, they additionally adopted a self-representation formulation (using U as the base).

(b) Noise measure.: Denoising aims to separate the noise from noisy data. However, the original data distribution may be still influenced after separation processing, as partial data without noise influence might be separated out as well during the process of separation. Meanwhile, as the noise distribution is usually discrete and scattered randomly, the noise influence on each example/example-label pair is different. That is, the noise contents of the example-label pairs are diverse, and it is clear that the higher the noise content, the more severely data distribution is destroyed. Therefore, to preserve the data distribution and better observe the noise content of each example/example-label pair, a good choice is to learn a noise distribution for a dataset, which can be used to calculate the noise content of each example/example-label pair. That is, the noise influence is quantized. Then, based on the extracted noise, the examples/example-label pairs that are less influenced by noise can be selected with high priority. Obviously, it can be useful to reduce the noise influence on the original data distribution.

For instance, having the noise matrix E obtained from Equation (43), Wu et al. [93] further quantify the noise QN_j of each unlabeled example x_j, as follows:

$$QN_j = \lim_{p \to \infty}\left(\sum_{s=1}^{d}\left(|E_{j,s}|\right)^p\right)^{\frac{1}{p}}, \tag{44}$$

where $E_{j,s}$ denotes the sth element of the noise vector E_j of example x_j.

<u>F. Expected Error Reduction Metric.:</u> Definition 15. Expected error reduction metric: Given an unlabeled set U and a known label space A, an expected error reduction metric is a function Err used to estimate the reduction of expected error after labeling an unlabeled example/example-label pair:

$$Err\colon\begin{cases} U \mapsto \mathscr{R} & \text{example-based} \\ U \times A \mapsto \mathscr{R} & \text{example-label-based} \end{cases}. \tag{45}$$

Since achieving a high-performance classifier is the ultimate objective of a multi-label active learning scheme, an unlabeled example/example-label pair that can directly enhance the generalization ability of the current classifier is the best choice and should be selected to label preferentially. Thus, enhancing the generalization ability can also be regarded as reducing the expected error [2], which leads to an expected error reduction metric.

Taking an example-based active learning method for example, expected error reduction sampling has three main steps. First, each example x in the unlabeled dataset U is temporarily added to the training set as a candidate example with each potential label to retrain a classifier. Second, the expected classification error of this unlabeled example is calculated from the classification error of each retrained classifier. Finally, after obtaining the expected classification error of each unlabeled example, the unlabeled example with the greatest expected error reduction is selected to query its label and then truly added into the training set.

There are many studies whose sampling strategies are designed based on the expected error reduction metric. For instance, a more direct way is to calculate the decrement of the expected loss between two contiguous learning cycles. Li et al. [54] named it as Mean Max Loss (MML), which is defined as follows:

$$\left\{x_j^*\right\}^t = \underset{\left\{x_j\right\}_s^t \subset U^t}{\arg\max} \sum_{x_j \in \left\{x_j\right\}_s^t} \frac{1}{\left|\hat{Y}\right|} \left\{ \sum_{y_{j,s} \in \hat{Y}, s = 1}^{\left|\hat{Y}\right|} \sum_{k=1}^{n_m} \max\left[\left(1 - m_{sk} f_k\left(x_j\right)\right), 0\right] \right\}, \tag{46}$$

where $|\hat{Y}|$ is the number of the predicted labels in the predicted label set $\hat{Y}$ of the example x_j, m_{sk} is a component of a coding matrix M [80] that is a $n_m \times n_m$ matrix with diagonal components 1 and others −1, $f_k(x_j)$ represents the output on the kth binary SVM classifier of x_j, t is the tth iteration, and $\left\{x_j\right\}_s^t$ is the corresponding selected example set from the unlabeled example set U in the tth iteration.

However, this approach is troubled by its time-consuming process. To simplify the computation, Li et al. [54] further considered a special case of MML that only computes the loss value on the most certain predicted label for an example. It is named Max Loss (ML), which can be represented as follows:

$$\left\{x_j^*\right\}^t = \underset{\left\{x_j\right\}_s^t \subset U^t}{\arg\max} \sum_{x_j \in \left\{x_j\right\}_s^t} \sum_{k=1}^{n_m} \max\left[\left(1 - m_{sk} f_k\left(x_j\right)\right), 0\right], \tag{47}$$

where s is the label index of the most certain predicted label $y_{j,s}$ of x_j.

Whereas Yang et al. [102] evaluated the loss function from the perspective of model depending on the size of version space [80]. A heuristics idea [79] is applied to approximate the new version space by the mapping between current SVM classification margin and the size of this new version space. The loss reduction $Loss(\cdot)$ of each binary classifier $f_{L+\left(x_j, y_{j,k}\right)}^k$ after the addition of the example-label pair $(x_j, y_{j,k})$ can be approximated as follows:

$$\frac{Loss\left(f_{L+\left(x_j, y_{j,k}\right)}^k\right)}{Loss\left(f_L^k\right)} \approx \frac{V_{L+\left(x_j, y_{j,k}\right)}^k}{V_L^k} \approx \frac{1 + y_{j,k} f_L^k\left(x_j\right)}{2}, \tag{48}$$

where $f^{k}_{L+(x_j, y_{j,k})}$ is a classifier trained from the training set L corresponding to the kth label and $V^{k}_{L+(x_j, y_{j,k})}$ is the size of its version space. Accordingly, the loss reduction is denoted as follows:

$$\begin{aligned} Loss\left(f_{L+(x_j, y_{j,k})}\right) - Loss(f_L) &= \sum_{k=1}^{n_m}\left(Loss\left(f^{k}_{L+(x_j, y_{j,k})}\right) - Loss\left(f^{k}_{L}\right)\right) \\ &= \sum_{k=1}^{n_m}\left(Loss\left(f^{k}_{L+(x_j, y_{j,k})}\right)\cdot\left(1 - \frac{Loss\left(f^{k}_{L}\right)}{Loss\left(f^{k}_{L+(x_j, y_{j,k})}\right)}\right)\right) \\ &\propto \sum_{k=1}^{n_m}\left(\frac{1 - y_{j,k} f^{k}_{L+(x_j, y_{j,k})}(x_j)}{2}\right), \end{aligned} \tag{49}$$

where f^{k}_{L} is the classifier before adding the example-label pair $(x_j, y_{j,k})$.

Some existing loss functions are also useful to measure the expected error reduction, such as Hamming-Loss [60] and Bayesian error bound [33, 63, 64, 78]. Here, Hamming-Loss is defined as follows:

$$HamLoss = \frac{1}{n_u}\sum_{j=1}^{n_u}\frac{\left|L(x_j)\Delta\hat{Y}_j\right|}{n_m}, \tag{50}$$

where $\hat{Y}_j = f(x_j)$ represents the predicted label set of the example x_j from a classifier Θ. The smaller value of the Hamming-Loss indicates a better classification performance. Bayesian error bound can be defined as follows:

$$Err(x_j, y_{j,k}) = \varepsilon_a - \varepsilon_b, \tag{51}$$

where ε_a is the expected Bayesian classification error of each example-label pair $(x_j, y_{j,k})$ over all the unlabeled examples in U before selection and defined as follows:

$$\varepsilon_a = \frac{1}{n_u}\sum_{x_j \in U,\, y_{j,k} \in LD(x_j)} \varepsilon(f \mid y_{j,k}, x_j), \tag{52}$$

where f is the classifier and ε_b is the expected Bayesian classification error after selection, which is defined as follows:

$$\begin{aligned} \varepsilon_b = \frac{1}{n_u}\Bigg\{ &\sum_{x_p \in U - x_j,\, y_{j,k} \in LD(x_j)} \varepsilon(f \mid y_{j,k}, x_p) \\ - &\sum_{x_p \in U - x_j,\, y_{j,s} \in LD(x_j) + y_{j,k}} \varepsilon(f \mid y_{j,s}, x_p)\Bigg\}, \end{aligned} \tag{53}$$

where $LD(x_j) + y_{j,k}$ is the augmented training set.

In addition, two hinge losses were adopted by Li and Guo [52], which are related to the predicted separation margin on each unlabeled example, and their summation is considered as the error. The approximate generalization is shown in Equation (4).

3.2 Annotation

In this subsection, different annotation methods are illustrated. There are two main annotation ways. One is manual annotation done by domain experts, and the other is automatic annotation. Figure 6 shows the integration of different annotation methods into active learning pipeline. Here, high-informativeness examples/example-label pairs will be selected to query their true labels.

3.2.1 Domain Experts.—When using active learning, the most informative examples/example-label pairs in the unlabeled pool are iteratively selected for annotation and then used to augment the training set. Through iterations, errors might be amplified if the annotation is not correct. When non-experts are asked to conduct annotations, they might be subjective and lack relevant experience. On the contrary, the labeling results from domain experts might be more reliable and precise, since they have rich professional knowledge and labeling experience. To ensure the accuracy of annotation, domain experts are usually hired in most studies.

When using manual labeling, it is important to evaluate the labeling reliability of the experts, specifically:

A. Intra-expert reliability: the measure of how consistent the labeling of the same example is when performed by the same expert at different occasions.

B. Inter-expert reliability: the measure of how consistent the labeling of the same example is when performed by different experts.

Selecting experienced professionals to perform the labeling improves intra-expert [6] and inter-expert [87] reliability. In addition, creating a detailed labeling protocol increases both intra- and inter-expert reliability measures [19].

Comparing with labeling by non-experts, the cost of domain experts is expensive. Especially, when the number of labels of each example is larger, the cost further increases. In the era of big data, we need to build models for more complicated problems, which usually need more labeled data to train a good model. Thus, a certain level of automatic annotation is necessary. In the next subsection, automatic annotation is described in detail.

3.2.2 Voting Committee.—As domain experts are costly and annotation workload is too heavy to be handled by domain experts, automatic annotation is adopted in many multi-label active learning studies, which can greatly reduce the labeling cost. We can understand the process of automatic annotation via voting committee from Figure 6. Specifically, in each iteration of a voting committee approach, the candidate examples/example-label pairs that are selected by a selection strategy are divided into two parts following an allocation criterion. One part of the candidate examples/example-label pairs are labeled by domain experts as an auxiliary for automatic annotation, while the other part is automatically labeled by a voting committee. The allocation criteria and the approaches of constructing the voting committee are the main components of automatic annotation. The latter is of great importance. Depending on various modeling approaches, existing approaches of constructing the voting committee are separated into two categories: multi-source information fusion based and multi-classifier based.

A. Multi-Source Information Fusion.: There exists much potential information that can be obtained during the process of active learning, some of which were described before. The potential information can be used to construct the voting committee.

With a goal to effectively utilize the remaining unlabeled data after some unlabeled data have been selected for manual labeling (contrast to automatic annotation), Tur et al. [86] proposed an automatic annotation method using the uncertainty. They claimed that the unlabeled example with the lowest uncertainty can be predicted with a high confidence, which can be used to augment the training set directly. To maximize the usage of manually labeled examples, they first retrained a classifier using the manually labeled examples and then made predictions for the examples in the unlabeled data pool using the classifier. Furthermore, they set a threshold to control the automatic process. Only the examples whose uncertainties are lower than the threshold will be selected to conduct automatic annotation.

For the sake of the improvement of the accuracy for automatic annotation, Wu et al. [95] proposed a SLMAL method, which captured three aspects of information, not only uncertainty but also nearest-neighbor and label correlation, to automatically annotate selected example-label pairs. After the most informative example-label pair is selected, they first determined whether this example-label pair can be automatically annotated correctly. If not, then it is sent to be labeled by domain experts.

B. Multi-Classifier.: Building multiple classifiers is another way to construct the voting committee for automatic annotation. Basically, it integrates the classification results from multiple classifiers to predict the labels for some unlabeled examples.

As an example, Wu et al. [97] constructed a committee using the classifiers trained for each label. Specifically, they utilized the conditional dependence information to determine a correlated $Label$ set for each label. Supposing that the correlated label set of $y_{j,k}$ of the example x_j is $\{Label_1, Label_2, \ldots, Label_t\}$, t classifers $\{F_1, F_2, \ldots, F_t\}$ are built on the original feature space augmented by each label in the correlated label set of $(x_j, y_{j,k})$. Then, they constructed a voting committee for example-label $(x_j, y_{j,k})$. The label $y_{j,k}$ of the example x_j will be automatically annotated if a high voting consistency exists. Otherwise, the example-label pair remains for human experts.

3.3 Discussion

In this section, we discussed existing multi-label active learning methods for image classification, focusing on two main sub-tasks, sampling and annotation. We discussed the advantages and disadvantages of each multi-label active learning algorithm. The summary of the state-of-the-art multi-label active learning methods is presented in Table 2. A check symbol (×) indicates that a multi-label active learning method has the feature(s) shown in columns.

From Table 2, we can draw the following conclusions. First, the selection granularity based on example-label pairs is becoming a popular research topic. Comparing to example-based sampling, it can dramatically reduce the annotation cost. Second, recent works are focused on designing sampling strategies, in which many different information sources are combined. The aim is to make use of various types of information from the feature and the label spaces such as uncertainty, label correlation, representativeness, diversity, noise content, and expected error reduction. Among the six information measures, expected error reduction has the highest computational complexity. There still exists additional potential information that can be mined from the example and the label dimensions that could be used to develop new metrics for designing sampling strategies. This has a lot of potential for future research. Third, combining domain expert and automatic labeling is being investigated to further reduce the annotation cost [30, 95–97]. The most important issue to consider here is how to guarantee the annotation accuracy.

From sampling granularity, the most commonly used are the example-based and example-label-based, as summarized in Section 3.1.1. A lot of effort has been put into example-label-based multi-label active learning in recent years, as it allows to simultaneously consider the example and the label dimensions. Our team has been studying the example-label-based sampling methods for multi-label active learning since 2014, and we have been focusing on the following aspects: label correlation, noise content, and automatically labeling. For the detailed description and experiments on these topics, please refer to our team’s multiple publications [30, 38–40, 91, 93–98, 105, 106, 116]). In addition, we also summarized available open source libraries and commonly used evaluation metrics as below.

Available Open Source Libraries.—There are several open source libraries for multi-label learning. MULAN [84] offers a plethora of state-of-the-art algorithms for multi-label classification and label ranking. In addition, it provides an evaluation framework that computes a large variety of multi-label evaluation measures through hold-out evaluation and cross-validation. It is written in Java and built on top of Weka [31], which is one of the most popular libraries for supervised learning. JCLAL [67] is another machine learning open source software for active learning, which uses both the WEKA and the MULAN libraries.

Evaluation Metrics.—Any binary evaluation metric can be used, commonly including precision, recall, accuracy, and F1-score [25]. For F1-score, two different approaches can be used: the *macro* approach, which computes one metric for each label and then the values are averaged over all categories, and the *micro* approach, which considers predictions from all instances together and then calculates the measure across all labels. When several active learning strategies are assessed, the visual comparison of learning curves may not be the best choice to conclude whether a strategy is significantly better than others. Therefore, Reyes et al. [65] proposed two comparison approaches based on the use of non-parametric statistical tests to statistically compare active learning strategies.

4 FUTURE PROMISE

The effectiveness of active learning in multi-label image classification is demonstrated by extensive efforts. However, there are still many open challenges that need to be resolved. In this section, we will discuss six worth-studying issues of multi-label active learning for image classification in terms of example dimension, label dimension, annotation, and application extension, as shown in Figure 7. As for the problems that have already been studied, we try to highlight existing difficulties. We hope these highlights can guide researchers to develop better solutions. As for brand new problems, we will give some suggestions for future studies.

4.1 Example Dimension

4.1.1 Weak-Label Problem.—Many previous studies on example-based multi-label active learning have a basic assumption. That is, all the labels of each selected example in a training set are given. However, this assumption is hardly established, since obtaining all the labels is costly and practically difficult. In practical applications, only partial labels of examples are usually provided. Under such a situation, the existence of a label does mean that the example is associated with this label, but the absence of a label cannot imply that this label does not belong to this example. This phenomenon is known as a weak-label problem [8, 12, 77, 97, 104, 108, 116].

Some existing studies solved the weak-label problem by decomposing it into a series of PU-learning (positive and unlabeled data learning) problems [18, 23, 53, 56] corresponding to each label, of which all labels are independent. However, as we discussed, there exist intrinsic relationships among labels, which can be used to improve the classifier's performance. Such a simple decomposition can reduce the impact of the correlations among labels. To solve the weak-label problem, there are two possible ways as described below.

One way is to automatically assign labels to fill in an incomplete training set and capture more useful information about the label distribution. Based on the complete training set, a more correct and comprehensive label correlation can be captured using a statistical method. It is clear that the core of this way is how to handle missing labels. A common solution adopted in the models SMSE2 [12], WELL [77], and MLR-GL [8] is to treat the missing labels as negative labels to complement the incomplete training set. This assumption can be made for multi-label active learning when most labels of each example in a dataset are negative. However, this is not true in most scenarios. Considering that there is an intrinsic mapping relation between the example space and the label space of a dataset, Wu et al. [91] proposed to capture the mapping structure to help predict the missing labels of each example in the training set by its vector representation in the feature space.

Another way is to design a suitable sampling strategy to explore a comprehensive label correlation. Since it is difficult to directly obtain the global unconditional label dependence in the weak-label context, Zhao et al. [97, 116] intended to explore the conditional label dependence with the aid of input features. However, in their studies, the way of calculating conditional label dependence needs plenty of computations. Besides, Yang et al. [104] proposed a MIML with weak-label (MIMLwel) approach, which takes the label relation into account under an assumption that some common examples are shared by some highly relevant labels. By making use of the mapping relation among the examples and their labels, Guo et al. [30] further proposed to capture the label correlation based on the label distribution buried in the mapping structure.

4.1.2 Example Noise Problem.—

Example noise is a universal phenomenon in real-word applications [49]. There are many possible reasons resulting in the noise problem, e.g., being contaminated during the process of data collection and data storage. Existing multi-label active learning studies for image classification always neglect possible noise in multi-label datasets. Noisy data misleads the estimation of data distribution and subsequently deteriorates the performance of a model built from the data. Therefore, how to reduce the interference of the noise in the training set is a much more important problem, which needs to be considered.

An applicable solution is to separate out the noise and obtain the original data. It has been widely used in other fields. For example, Meng et al. [58] proposed a low-rank matrix factorization (LRMF) method to estimate subspaces with an unknown noise distribution, which is useful to separate out the noise. In addition, aiming at adapting to more complex noise, Cao et al. [9] regarded noise as Mixture of Exponential Power (MoEP) distributions and proposed a penalized MoEP model by combining the penalized likelihood method with MoEP distributions. Based on the penalized MoEP model, a new LRMF model was developed. Inspired by this, Wu et al. [91] introduced a low-rank matrix recovery method in the active learning framework to separate out the noise from the dataset and extracted a low-rank representation to replace the noisy data for model training.

Another better choice is to design a specific sampling strategy for selecting the most dependable examples that contain less noise. Image noise would result in the implicit deviations of sample data from their actual values, and the influences on the examples are different because of the randomness of the noise. Clearly, the examples with less noise would contribute more to the model training than those that are severely influenced. Based on this idea, Lee et al. [49] emphasized more on reliable examples that contain less noise while learning a model. A novel matrix factorization model was designed for modeling the deviation of sample data. Also, Wu et al. [93] incorporated the noise level measure into the sampling strategy that quantifies the noise content of each example in the unlabeled data pool.

Until now, only few works [91, 93] focus on the example noise issue of multi-label active learning. These works conducted the exploratory studies on how to reduce the influence of example noise for constructing an effective training set. Furthermore, due to the complexity of example noise, e.g., various types of noise (Gaussian noise [59], Laplacian noise [101], and so on), the influence of example noise on the sampling strategy needs to be further studied. A better noise measure would be a main research direction for this issue.

4.2 Label Dimension

4.2.1 Multi-Instance Problem.—

In real-word applications, the semantic information of an image is usually ambiguous. It leads to the fact that an image can not only be described by a number of labels but also be associated with a number of instances simultaneously [110, 119–121]. It is known as a multi-instance multi-label problem, which is first formalized by Zhou et al. [120].

To solve a multi-instance multi-label active learning task, an admissible solution is to use multi-label learning as the bridge, which is proposed in References [120, 121]. More precisely, a multi-instance multi-label learning problem can be decomposed into a set of multi-label learning problems by extending existing multi-label active learning methods. However, if we directly decompose the multi-instance multi-label learning problems, then it might lose sight of the correlation among instances.

Whereas Zha et al. [110] designed an integrated framework depending on the observation that each individual label of one image is actually related to some local regions rather than an entire image. Regarding each region as an independent part, it can be viewed as an instance respectively. Then, the image can be seen as a bag of instances. Accordingly, the multi-instance multi-label problem can be converted into a set of single-instance problems. Then, corresponding approaches in multi-instance learning can be adopted to deal with the multi-instance multi-label issue.

4.2.2 Label Tree.—As discussed, the performance of a classifier can be improved with the help of label correlation exploration. There is a special label relationship that can be demonstrated in a hierarchical structure, as there exist inclusion relations among labels.

As for two types of label correlation exploration in existing studies as we have mentioned in Section 3.1.2, the discussed relationship is usually the relationship between a label pair, which neglects the label correlation among multiple labels. The tree structure is useful to describe the label correlations among multiple labels. Regarding each label as one node of a tree, relevant label pairs can be represented as two nodes connecting by one branch. Specifically, in the same subtree, there is an inclusion relation between a parent node and a child node, and the relationship between sibling nodes is also relevant. The relationship between nodes in different subtrees is independent. After constructing a label tree to represent the hierarchical structure among all the labels in the dataset, only the labels on the leaf nodes need to be annotated. It is easy to automatically annotate the labels on their upper nodes. Therefore, constructing a label tree is a very useful way to represent the label correlation. A following-up question is how to construct a label tree.

The approach used to explore conditional label dependence in References [97, 98, 116] is a reasonable way. First, they obtained the conditional dependence among each label pair and ranked these label pairs in terms of the degree of dependence. Then, a threshold is set to select a set of high-dependence label pairs. Different label pairs that one of their component labels is the same can be integrated. For instance, if a label a is conditional dependent on a label b (i.e., $a \rightarrow b$) and the label a is also conditional dependent on a label c (i.e., $a \rightarrow c$), then it can be integrated as $a \rightarrow \{b, c\}$. After integrating all the possible suitable label pairs, a label tree is constructed.

4.3 Annotation

4.3.1 Automatic Annotation.—As the process of active learning is iterative, it is more likely to amplify the annotation errors occurred in previous iterations. That is why most multi-label active learning studies for image classification request labels from domain experts to ensure the accuracy of annotation. However, it is costly to hire domain experts to conduct annotation (as mentioned). Inspired by semi-supervised learning, it can train classifiers exploiting a limited number of labeled examples and a large number of unlabeled examples [3, 4, 12, 75, 89, 118, 122], incorporating automatical annotation into multi-label active learning attracts much more attention. Apparently, the most important issue of automatic annotation is its annotation accuracy.

Inspired by Query by Committee (QBC) [70], constructing a voting committee is a good choice to conduct the automatic annotation. As discussed in Section 3.2.2, two schemes can be used to construct a voting committee. One is a multi-source information-based scheme, and another is a multi-classifier-based scheme. The major challenge of the multi-source information-based scheme is how to explore different types of potential information and utilize them. The major challenge of the multi-classifier-based scheme is how to construct an effective training set for each classifier and choose proper learning algorithms to build the classifiers from the training set.

4.3.2 Crowdsourcing.—As mentioned, annotation by domain experts is costly. Moreover, with the increment of labeling time, the labeling precision could decrease because of fatigue. These factors need to be further discussed and considered.

One reliable solution, also an important research direction, is to incorporate crowd computing into active learning, which is well known as crowdsourcing. Crowdsourcing is first introduced in Reference [32]. Basically, crowdsourcing means that a group of non-experts (workers) work on the tasks that cannot be easily done by individuals. Comparing to annotating by domain experts, the application of crowdsourcing in multi-label active learning is able to save the cost, due to the low cost of non-experts.

The high-informativeness examples/example-label pairs are handed out to a batch of network users for annotation. Different from annotating by domain experts, crowdsourcing allows several workers to repeatedly annotate an unlabeled example several times, i.e., repeated labeling [71]. After obtaining the repeated annotating results of the same example, a consensus algorithm is applied to estimate its ground truth. Then, the estimated ground truth is treated as the label for this example to augment the training set. The advantage of crowdsourcing is that a plenty of cheap labels can be obtained in a short time, while its disadvantage is that the obtained labels may have errors or omissions because of different labeling abilities of labelers. Thus, how to integrate network users’ annotations effectively is an important issue on this topic. To improve data quality, Sheng et al. [71, 72] proposed several intelligent repeated-labeling strategies. With some recent advent of inexperience and scalable online annotator tools (e.g. Amazon’s Mechanical Turk and Crowdflower), Donmez et al. [17] presented a thresholding mechanism (IEThresh) to estimate a confidence interval to represent the reliability of each worker and select the one with the highest estimated annotation accuracy. Zhang et al. [113] proposed a bilayer collaborative clustering method for the label aggregation in crowdsourcing.

4.4 Application Extension

4.4.1 Semantic Segmentation.—Some other computer vision tasks can also be included in image classification such as semantic image segmentation. The goal of semantic image segmentation is to label each pixel of an image with a corresponding class. Recent advances in semantic segmentation have enabled their application to medical image segmentation [55]. Convolutional Neural Network– (CNN) based deep learning algorithms became a natural choice for medical image segmentation tasks. Medical imaging data varies between different imaging modalities and protocols. Therefore, each new segmentation task usually needs a new dataset. Creating a training dataset involves manual segmentation of three-dimensional imaging volumes by experienced clinical professionals. In addition, each voxel may correspond to multiple labels (*multi-label semantic segmentation*). Therefore, creating such training dataset requires a great deal of effort and cost.

A deep active learning framework, which combines fully convolutional neural network and active learning, could be developed to significantly reduce annotation effort. Active learning would help decide which images should be annotated to achieve the best performance with limited time and budget. A few recent works [74, 103] have been presented to determine the most uncertain and representative examples for annotation. A new deep active learning framework that fuses multi-source information from subject data and the label space to select the most informative subject data for annotation would help us achieve optimal semantic segmentation results with limited resources for manual segmentation of training data.

4.4.2 Active Feature Selection.—An important aspect of this problem has usually been ignored. Base classifiers are not specifically designed for active learning; they often require ample labeled data [5]. When the number of features is large and the training data are limited, it is difficult to get reliable estimates of model parameters. This could create an inaccurate model and adversely affect the whole active learning process.

For example, radiomics is the high-throughput extraction and analysis of numerous features from medical images. It represents a highly promising approach for characterizing tumor phenotypes, providing an unprecedented opportunity to support and improve personalized clinical decision-making [26, 92]. However, the reliable and efficient usage of radiomic features for an accurate cancer outcome prediction remains very challenging due to the uncertainty and the redundancy of radiomic features extracted from medical images.

Feature selection refers to the study of algorithms selecting an optimal subset from the input feature set. Feature selection algorithms are widely used in machine learning to reduce the feature space representing given data samples. Inspired by the idea of multi-label active learning, we can develop an effective sampling strategy to select the most informative radiomic features from an unselected radiomics feature pool and then construct an effective training set with strong discriminative radiomic features. Using active learning on simultaneous example and feature selection combines faster learning with a smaller feature space dimensionality.

In recent years, a new direction in cancer research, radiogenomics [76], has emerged that aims to discover clinically actionable association between high-throughput features extracted from medical images and high-throughput genomic data. Radiogenomics studies have focused primarily on investigating selection of salient features to predict a pathological stage [107]. Using active learning on simultaneous example and feature selection for radiogenomics is another promising research direction.

In summary, we discussed future promises of multi-label active learning from four aspects, i.e., example dimension, label dimension, annotation, and application extension. A key area of future research should focus on improving sampling strategies to construct an effective training set for various application scenarios.

5 CONCLUSION

In this work, we provided a systematic overview of multi-label active learning for image classification. Existing multi-label active learning algorithms have been categorized into two top groups, namely sampling and annotation. The corresponding representative works in each group have been discussed in detail, and recommendations for future improvements have also been discussed. We then summarized major challenges for multi-label active learning, together with the promising trends that might be developed in the future. However, this survey is not exhaustive and the research in this field is far from being over. We expect the multi-label active learning for image classification to become more widespread due to the key advantage of utilizing abundant unlabeled images with limited resources.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under grant No. 61402311, 91846104, 61728205, 61876117, 61876217, 61603186, and the U.S. National Science Foundation (IIS-1115417).

REFERENCES

[1]. Agrawal Rakesh, Imielinski Tomasz, and Swami Arun. 1993. Mining association rules between sets of items in large databases. In Proceedings of theACM International Conference on Management of Data (SIGMOD'93). 207–216.

[2]. Allwein Erin L., Schapire Robert E., and Singer Yoram. 2000. Reducing multiclass to binary: A unifying approach for margin classifiers. In Proceedings of the17th International Conference on Machine Learning (ICML'00). 9–16.

[3]. Amini Massih R. and Gallinari Patrick. 2005. Semi-supervised learning with an imperfect supervisor. Knowl. Inf. Syst 13, 1 (2005), 1–42.

[4]. Angluin Dana. 1988. Queries and concept learning. Mach. Learn2, 4 (1988), 319–342.

[5]. Bilgic Mustafa. 2012. Combining active learning and dynamic dimensionality reduction. In Proceedings of the2012 SIAM International Conference on Data Mining. SIAM, 696–707.

[6]. Bø Hans Kristian, Solheim Ole, Jakola Asgeir Store, Kvistad Kjell-Arne, Reinertsen Ingerid, and Berntsen Erik Magnus. 2017. Intra-rater variability in low-grade glioma segmentation. J. Neuro-oncol131, 2 (2017), 393–402.

[7]. Boutell Matthew R., Luo Jiebo, Shen Xipeng, and Brown Christopher M.. 2004. Learning multi-label scene classification. Pattern Recogn37, 9 (2004), 1757–1771.

[8]. Bucak Selcuk, Serhat Rong Jin, and Jain Anil K.. 2011. Multi-label learning with incomplete class assignments. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR’11). 2801–2808.

[9]. Cao Xiangyong, Chen Yang, Zhao Qian, Meng Deyu, Wang Yao, Wang Dong, and Xu Zongben. 2016. Low-rank matrix factorization under general mixture noise distributions. In Proceedings of theIEEE International Conference on Computer Vision (ICCV’16). 1493–1501.

[10]. Chakraborty Shayok, Balasubramanian Vineeth, and Panchanathan Sethuraman. 2011. Optimal batch selection for active learning in multi-label classification. In Proceedings of the19th ACM International Conference on Multimedia. ACM, 1413–1416.

[11]. Chen Beijing, Shu Huazhong, Coatrieux Gouenou, Chen Gang, Sun Xingming, and Coatrieux Jean Louis. 2015. Color image analysis by quaternion-type moments. J. Math. Imag. Vis51, 1 (2015), 124–144.

[12]. Chen Gang, Song Yangqiu, Wang Fei, and Zhang Changshui. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In Proceedings of theSIAM International Conference on Data Mining (SDM’08). 410–419.

[13]. Myung Jin Choi Antonio Torralba, and Willsky Alan S.. 2012. A tree-based context model for object recognition. IEEE Trans. Pattern Anal. Mach. Intell34, 2 (2012), 240–252. [PubMed: 21670482]

[14]. Cohn David, Atlas Les, and Ladner Richard. 1994. Improving generalization with active learning. Mach. Learn15, 2 (1994), 201–221.

[15]. Cohn David. A., Ghahramani Zoubin, and Jordan Michael I.. 1996. Active learning with statistical models. J. Artif. Intell. Res4, 1 (1996), 705–712.

[16]. Dagan Ido and Engelson Sean P.. 1995. Committee-based sampling for training probabilistic classifiers. In Proceedings of the 12th International Conference on Machine Learning (ICML’95). 150–157.

[17]. Donmez Pinar, Carbonell Jaime G., and Schneider Jeff. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In Proceedings of theACM International Conference on Knowledge Discovery and Data Mining (SIGKDD’09). 259–268.

[18]. Elkan Charles and Noto Keith. 2008. Learning classifiers from only positive and unlabeled data. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD’08). 213–220.

[19]. Entis Jonathan J., Doerga Priya, Barrett Lisa Feldman, and Dickerson Bradford C.. 2012. A reliable protocol for the manual segmentation of the human amygdala and its subregions using ultra-high resolution MRI. Neuroimage60, 2 (2012), 1226–1235. [PubMed: 22245260]

[20]. Esuli Andrea and Sebastiani Fabrizio. 2009. Active learning strategies for multi-label text classification. In Proceedings of the European Conference on Information Retrieval (ECIR’09). 102–113.

[21]. Fu Yifan, Zhu Xingquan, and Li Bin. 2013. A survey on instance selection for active learning. Knowl. Inf. Syst35, 2 (2013), 249–283.

[22]. Fujii Atsushi, Tokunaga Takenobu, Inui Kentaro, and Tanaka Hozumi. 1998. Selective sampling for example-based word sense disambiguation. Comput. Linguist24, 4 (1998), 573–597.

[23]. Fung Gabriel Pui Cheong, Yu Jeffrey X., Lu Hongjun, and Yu Philip S.. 2006. Text classification without negative examples revisit. IEEE Trans. Knowl. Data Eng18, 1 (2006), 6–20.

[24]. Gao Nengneng, Huang Sheng-Jun, and Chen Songcan. 2016. Multi-label active learning by model guided distribution matching. Front. Comput. Sci10, 5 (2016), 845–855.

[25]. Gibaja Eva and Ventura Sebastián. 2015. A tutorial on multilabel learning. ACM Comput. Surv 47, 3 (2015), 52.

[26]. Gillies Robert J., Kinahan Paul E., and Hricak Hedvig. 2015. Radiomics: Images are more than pictures, they are data. Radiology278, 2 (2015), 563–577. [PubMed: 26579733]

[27]. Gu Bin and Sheng Victor S.. 2016. A robust regularization path algorithm for v-support vector classification. IEEE Trans. Neur. Netw. Learn. Syst 1, 99 (2016), 1–8.

[28]. Gu Bin, Sheng Victor S., and Li Shuo. 2015. Bi-parameter space partition for cost-sensitive SVM. In Proceedings of theInternational Conference on Artificial Intelligence (AAAI’15). 3532–3539.

[29]. Gu Bin, Sun Xingming, and Sheng Victor. S.. 2017. Structural minimax probability machine. IEEE Trans. Neur. Netw. Learn. Syst28, 7 (2017), 1646–1656.

[30]. Guo Anqian, Wu Jian, Sheng Victor S., Zhao Pengpeng, and Cui Zhiming. 2017. Multi-label active learning with low-rank mapping for image classification. In Proceedings of theIEEE International Conference on Multimedia and Expo (ICME’17). 259–264.

[31]. Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter, and Witten Ian H.. 2009. The WEKA data mining software: An update. ACM SIGKDD Explor. Newslett11, 1 (2009), 10–18.

[32]. Howe Jeff. 2006. The rise of crowdsourcing. Wired Mag14, 6 (2006), 1–4.

[33]. Hua Xian Sheng and Qi Guo Jun. 2008. Online multi-label active annotation: Towards large-scale content-based video search. In Proceedings of the ACM International Conference on Multimedia (ACM MM’08). 141–150.

[34]. Huang Shengjun, Jin Rong, and Zhou Zhihua. 2010. Active learning by querying informative and representative examples. In Proceedings of theInternational Conference on Neural Information Processing Systems. 892–900.

[35]. Huang Shengjun, Jin Rong, and Zhou Zhihua. 2014. Active learning by querying informative and representative examples. IEEE Trans. Pattern Anal. Mach. Intell36, 10 (2014), 1936–49. [PubMed: 26352626]

[36]. Huang Sheng Jun, Chen Songcan, and Zhou Zhihua. 2015. Multi-label active learning: Query type matters. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI’15), 946–952.

[37]. Yu Gang Jiang Qi Dai, Wang Jun, Chong Wah Ngo Xiangyang Xue, and Chang Shih Fu. 2012. Fast semantic diffusion for large-scale context-based image and video annotation. IEEE Trans. Image Process21, 6 (2012), 3080–3091. [PubMed: 22345543]

[38]. Jiao Yang. 2015. Active Learning-Based Multi-Label Image Classification. Soochow University in China.

[39]. Jiao Yang, Zhao Pengpeng, Wu Jian, Shi Yujie, and Cui Zhiming. 2014. A multicriterion query-based batch mode active learning technique. In Foundations of Intelligent Systems. 669–680.

[40]. Jiao Yang, Zhao Pengpeng, Wu Jian, Xian Xuefeng, Xu Haihui, and Cui Zhiming. 2014. Active multi-label learning with optimal label subset selection. In Proceedings of theInternational Conference on Advanced Data Mining and Applications (ADMA’14). 523–534.

[41]. Joshi Ajay J., Porikli Fatih, and Papanikolopoulos Nikolaos. 2009. Multi-class active learning for image classification. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR’09). 2372–2379.

[42]. Kapoor Ashish, Grauman Kristen, Urtasun Raquel, and Darrell Trevor. 2007. Active learning with gaussian processes for object categorization. In Proceedings of theIEEE International Conference on Computer Vision (ICCV’07), Vol. 88. 1–8.

[43]. Kawewong Aram, Pimpup Rapeeporn, and Hasegawa Osamu. 2013. Incremental learning framework for indoor scene recognition. In Proceedings of the27th AAAI Conference on Artificial Intelligence (AAAI’13). 496–502.

[44]. Kemp Freda. 2003. Applied multiple regression/correlation analysis for the behavioral sciences. J. Roy. Stat. Soc52, 4 (2003), 691–691.

[45]. King Ross D., Rowland Jem, Oliver Stephen G., Young Michael, Aubrey Wayne, Byrne Emma, Liakata Maria, Markham Magdalena, Pir Pinar, and Soldatova Larisa N.. 2009. The automation of science. Science324, 5923 (2009), 85–89. [PubMed: 19342587]

[46]. King Ross D., Whelan Kenneth E., Jones Ffion M., Reiser Philip G. K., Bryant Christopher H., Muggleton Stephen H., Kell Douglas B., and Oliver Stephen G.. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature427, 6971 (2004), 247–252. [PubMed: 14724639]

[47]. Krishnamurthy V. 2002. Algorithms for optimal scheduling and management of hidden Markov model sensors. IEEE Trans. Sign. Process50, 6 (2002), 1382–1397.

[48]. Lang Kenneth and Baum Eric. B.. 1992. Query learning can work poorly when a human oracle is used. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN’92).

[49]. Lee Guang He, Yang Shao Wen, and Lin Shou De. 2016. Toward implicit sample noise modeling: Deviation-driven matrix factorization. arXiv preprint arXiv:1610.09274 (2016).

[50]. Lewis David D. and Catlett Jason. 1994. Heterogenous uncertainty sampling for supervised learning. In Proceedings of the International Conference on Machine Learning (ICML'94). 148–156.

[51]. Li Jian, Li Xiaolong, Yang Bin, and Sun Xingming. 2015. Segmentation-based image copy-move forgery detection scheme. IEEE Trans. Inf. Forens. Secur10, 3 (2015), 507–518.

[52]. Li Xin and Guo Yuhong. 2013. Active learning with multi-label SVM classification. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'13). 1479–1485.

[53]. Li Xiaoli and Liu Bing. 2003. Learning to classify texts using positive and unlabeled data. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'03). 587–592.

[54]. Li Xuchun, Wang Lei., and Sung Eric.. 2004. Multilabel SVM active learning for image classification. In Proceedings of the International Conference on Image Processing (ICIP'04), Vol. 4. 2207–2210.

[55]. Litjens Geert, Kooi Thijs, Bejnordi Babak Ehteshami, Setio Arnaud Arindra Adiyoso, Ciompi Francesco, Ghafoorian Mohsen, Van Der Laak Jeroen Awm, Van Ginneken Bram, and Sánchez Clara I.. 2017. A survey on deep learning in medical image analysis. Med. Image Anal42 (2017), 60–88. [PubMed: 28778026]

[56]. Liu B, Dai Y, Li X, and Lee WS. 2003. Building text classifiers using positive and unlabeled examples. In Proceedings of theIEEE International Conference on Data Mining (ICDM'03). 179–186.

[57]. Liu Guangcan, Lin Z, Yan S, Sun J, Yu Y, and Ma Y. 2013. Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell35, 1 (2013), 171–184. [PubMed: 22487984]

[58]. Meng Deyu and De La Torre Fernando. 2013. Robust matrix factorization with unknown noise. In Proceedings of the IEEE International Conference on Computer Vision (ICCV'13), 1337–1344.

[59]. Mitra Kaushik, Sheorey Sameer, and Chellappa Rama. 2010. Large-scale matrix factorization with missing data under additional constraints. Ophthal. Res40, 1 (2010), 35.

[60]. Nasierding Gulisong and Kouzani Abbas Z.. 2010. Empirical study of multi-label classification methods for image annotation and retrieval. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA'10). 617–622.

[61]. Reports of International Data Corporation. 2010. Retrieved from http://www.idc.com/prodserv/prodserv.jsp.

[62]. Olsson Fredrik. 2009. A Literature Survey of Active Machine Learning in the Context of Natural Language Processing. Swedish Institute of Computer Science. Technical Report T2009:06.

[63]. Qi Guojun, Hua Xiansheng, Rui Yong, Tang Jinhui, and Zhang Hongjiang. 2008. Two-dimensional active learning for image classification. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). 1–8.

[64]. Qi Guojun, Hua Xiansheng, Rui Yong, Tang Jinhui, and Zhang Hongjiang. 2009. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. IEEE Trans. Pattern Anal. Mach. Intell31, 10 (2009), 1880–1897. [PubMed: 19696456]

[65]. Reyes Oscar, Altalhi Abdulrahman H., and Ventura Sebastián. 2018. Statistical comparisons of active learning strategies over multiple datasets. Knowl.-Based Syst145 (2018), 274–288.

[66]. Reyes Oscar, Morell Carlos, and Ventura Sebastián. 2018. Effective active learning strategy for multi-label learning. Neurocomputing273 (2018), 494–508.

[67]. Reyes Oscar, Pérez Eduardo, Del Carmen Rodríguez-Hernández María, Fardoun Habib M., and Ventura Sebastián. 2016. JCLAL: A Java framework for active learning. J. Mach. Learn. Res17, 1 (2016), 3271–3275.

[68]. Reyes Oscar and Ventura Sebastián. 2018. Evolutionary strategy to perform batch-mode active learning on multi-label data. ACM Trans. Intell. Syst. Technol 9, 4 (2018), 46.

[69]. Settles Burr. 2010. Active Learning Literature Survey. Computer Science Technical Report1648, University of Wisconsin-Madison.

[70]. Sebastian Seung H, Opper Manfred, and Sompolinsky Haim. 1992. Query by committee. In Proceedings of theThe Annual Workshop on Computational Learning Theory (COLT'92). 287–294.

[71]. Sheng Victor S., Provost Foster, and Ipeirotis Panagiotis G.. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In Proceedings of theACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08). 614–622.

[72]. Sheng Victor S. and Zhang Jing. 2019. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19), Vol. 33. 9837–9843.

[73]. Singh Mohan and Curran Eoin. 2008. Active learning for multi-label image annotation. In Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science (AICS'08). 173–182.

[74]. Smailagic Asim, Costa Pedro, Hae Young Noh Devesh Walawalkar, Khandelwal Kartik, Galdran Adrian, Mirshekari Mostafa, Fagert Jonathon, Xu Susu, Zhang Pei, et al.2018. MedAL: Accurate and robust deep active learning for medical image analysis. In Proceedings of the2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA'18). IEEE, 481–488.

[75]. Stikic Maja, Van Laerhoven Kristof, and Schiele Bernt. 2008. Exploring semi-supervised and active learning for activity recognition. In Proceedings of theIEEE International Symposium on Wearable Computers. 81–88.

[76]. Story Michael D. and Durante Marco. 2018. Radiogenomics. Medical Physics 45, 11 (2018), e1111–e1122. [PubMed: 30421807]

[77]. Sun Yu Yin, Zhang Yin, and Zhou Zhi Hua. 2010. Multi-label learning with weak label. In Proceedings of the24th AAAI Conference on Artificial Intelligence (AAAI'10). 593–598.

[78]. Tang Jinhui, Zha Zhengun, Tao Dacheng, and Chua Tatseng. 2012. Semantic-gap-oriented active learning for multilabel image annotation. IEEE Trans. Image Process21, 4 (2012), 2354–2360. [PubMed: 22194245]

[79]. Tong Simon. 2001. Active Learning: Theory and Applications. Stanford University.

[80]. Tong Simon and Koller Daphne. 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res 2, 11 (2001), 45–66.

[81]. Tsoumakas Grigorios, Katakis Ioannis, and Taniar David. 2007. Multi-label classification: An overview. Int. J. Data Warehous. Min3, 3 (2007), 1–13.

[82]. Tsoumakas Grigorios, Katakis Ioannis, and Vlahavas Ioannis. 2008. Effective and efficient multilabel classification in domains with large number of labels. In Proceedings of theECML/PKDD Workshop on Mining Multidimensional Data (MMD'08). 30–44.

[83]. Tsoumakas Grigorios, Katakis Ioannis, and Vlahavas Ioannis. 2009. Mining multi-label data. In Data Mining and Knowledge Discovery Handbook. Springer, 667–685.

[84]. Tsoumakas Grigorios, Eleftherios Spyromitros-Xioufis Jozef Vilcek, and Vlahavas Ioannis. 2011. MULAN: A Java library for multi-label learning. J. Mach. Learn. Res12, 7 (2011), 2411–2414.

[85]. Tuia Devis, Volpi Michele, Copa Loris, Kanevski Mikhail, and Munoz-Mari Jordi. 2011. A survey of active learning algorithms for supervised remote sensing image classification. IEEE J. Select. Top. Sign. Process5, 3 (2011), 606–617.

[86]. Tur Gokhan, Hakkani-Tur Dilek, and Schapire Robert E.. 2005. Combining active and semi-supervised learning for spoken language understanding. Speech Commun45, 2 (2005), 171–186.

[87]. Visser M, Müller DMJ, van Duijn RJM, Smits M, Verburg N, Hendriks EJ, Nabuurs RJA, Bot JCJ, Eijgelaar RS, Witte M, et al.2019. Inter-rater agreement in glioma segmentations on longitudinal MRI. NeuroImage: Clin22 (2019), 101727. [PubMed: 30825711]

[88]. Wang Meng and Hua Xiansheng. 2011. Active learning in multimedia annotation and retrieval: A survey. ACM Trans. Intell. Syst. Technol 2, 2 (2011), 1–21.

[89]. Wei Zhihua, Wang Hanli, and Zhao Rui. 2013. Semi-supervised multi-label image classification based on nearest neighbor editing. Neurocomputing119 (2013), 462–468.

[90]. Wen Xuezhi, Shao Ling, Xue Yu, and Fang Wei. 2015. A rapid learning algorithm for vehicle classification. Inf. Sci295, 1 (2015), 395–406.

[91]. Wu Jian, Guo Anqian, Sheng Victor S., Zhao Pengpeng, Cui Zhiming, and Li Hua. 2017. Adaptive low-rank multi-label active learning for image classification. In Proceedings of theACM International Conference on Multimedia (ACM MM'17). 1336–1344.

[92]. Wu Jian, Lian Chunfeng, Ruan Su, Mazur Thomas R., Mutic Sasa, Anastasio Mark A., Grigsby Perry W., Vera Pierre, and Li Hua. 2019. Treatment outcome prediction for cancer patients based on radiomics and belief function theory. IEEE Trans. Radiat. Plasma Med. Sci3, 2 (2019), 216–224. [PubMed: 31903444]

[93]. Wu Jian, Ruan Su, Lian Chunfeng, Mutic Sasa, Anastasio Mark A., and Li Hua. 2018. Active learning with noise modeling for medical image annotation. In Proceedings of the2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI'18). IEEE, 298–301.

[94]. Wu Jian, Sheng Victor S., Zhang Jing, Zhao Pengpeng, and Cui Zhiming. 2014. Multi-label active learning for image classification. In Proceedings of theIEEE International Conference on Image Processing (ICIP'14). 5227–5231.

[95]. Wu Jian, Ye Chen, Sheng Victor S., Yao Yufeng, Zhao Pengpeng, and Cui Zhiming. 2015. Semi-automatic labeling with active learning for multi-label image classification. In Proceedings of thePacific Rim Conference on Multimedia (PCM'15). 473–482.

[96]. Wu Jian, Ye Chen, Sheng Victor S., Zhang Jing, Zhao Peng Peng, and Cui Zhiming. 2017. Active learning with label correlation exploration for multi-label image classification. IET Comput. Vis11, 7 (2017), 577–584.

[97]. Wu Jian, Zhao Shiquan, Sheng Victor S., Zhang Jing, Ye Chen, Zhao Peng Peng, and Cui Zhiming. 2017. Weak labeled active learning with conditional label dependence for multi-label image classification. IEEE Trans. Multimedia19, 6 (2017), 1156–1169.

[98]. Wu Jian, Zhao Shiquan, Sheng Victor S., Zhao Pengpeng, and Cui Zhiming. 2016. Multi-label active learning for image classification with asymmetrical conditional dependence. In Proceedings of theIEEE International Conference on Multimedia and Expo (ICME'16). 5227–5231.

[99]. Xia Zhihua, Wang Xinhui, Sun Xingming, Liu Quansheng, and Xiong Naixue. 2016. Steganalysis of LSB matching using differences between nonadjacent pixels. Multimedia Tools Appl75, 4 (2016), 1947–1962.

[100]. Xu Xin Shun, Jiang Yuan, Peng Liang, Xue Xiangyang, and Zhou Zhi Hua. 2011. Ensemble approach based on conditional random field for multi-label image and video annotation. In Proceedings of theACM International Conference on Multimedia (ACM MM'11). 1377–1380.

[101]. Yan Shuicheng. 2012. Practical low-rank matrix approximation under robust L1-norm. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR'12). 1410–1417.

[102]. Yang Bishan, Sun Jian Tao, Wang Tengjiao, and Chen Zheng. 2009. Effective multi-label active learning for text classification. In Proceedings of theACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'09). 917–926.

[103]. Yang Lin, Zhang Yizhe, Chen Jianxu, Zhang Siyuan, and Chen Danny Z.. 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In Proceedings of theInternational Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'17). Springer, 399–407.

[104]. Yang Shu Jun, Jiang Yuan, and Zhou Zhi Hua. 2013. Multi-instance multi-label learning with weak label. In Proceedings of theInternational Joint Conference on Artificial Intelligence (IJCAI'13). 1862–1868.

[105]. Ye Chen, Wu Jian, Sheng Victor S., and Zhao Pengpeng. 2015. Multi-label active learning with label correlation for image classification. In Proceedings of theIEEE International Conference on Image Processing (ICIP'15). 3437–3441.

[106]. Ye Chen, Wu Jian, Sheng Victor S., Zhao Shiquan, Zhao Pengpeng, and Cui Zhiming. 2015. Multi-label active learning with chi-square statistics for image classification. In Proceedings of theACM International Conference on Multimedia Retrieval (ICMR'15). 583–586.

[107]. Yoon Hong-Jun, Ramanathan Arvind, Alamudun Folami, and Tourassi Georgia. 2018. Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer. In Proceedings of

the14th International Workshop on Breast Imaging (IWBI'18), Vol. 10718. International Society for Optics and Photonics, 107181H.

[108]. Yu Guoxian, Zhang Guoji, Rangwala Huzefa, Domeniconi Carlotta, and Yu Zhiwen. 2012. Protein function prediction using weak-label learning. In Proceedings of theACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB'12). 202–209.

[109]. Yu Hwanjo. 2005. SVM selective sampling for ranking with application to data retrieval. In Proceedings of the11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'05). 354–363.

[110]. Zha Zheng Jun, Hua Xian Sheng, Mei Tao, Wang Jingdong, Qi Guo Jun, and Wang Zengfu. 2008. Joint multi-label multi-instance learning for image classification. In Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). 1–8.

[111]. Zhang Bang, Wang Yang, and Chen Fang. 2014. Multilabel image classification via high-order label correlation driven active learning. IEEE Trans. Image Process23, 3 (2014), 1430–1441. [PubMed: 24723538]

[112]. Zhang Bang, Wang Yang, and Wang Wei. 2012. Batch mode active learning for multi-label image classification with informative label correlation mining. In Proceedings of theIEEE Workshop on the Applications of Computer Vision (WACV'12). 401–407.

[113]. Zhang Jing, Sheng Victor S., and Wu Jian. 2019. Crowdsourced label aggregation using bilayer collaborative clustering. IEEE Trans. Neur. Netw. Learn. Syst30, 10 (2019), 3172–3185.

[114]. Zhang Xiaoyu, Cheng Jian, Xu Changsheng, Lu Hanqing, and Ma Songde. 2009. Multi-view multi-label active learning for image classification. In Proceedings of theIEEE International Conference on Multimedia and Expo (ICME'09). 258–261.

[115]. Zhang Yi. 2010. Multi-task active learning with output constraints. In Proceedings of the24th AAAI Conference on Artificial Intelligence (AAAI'10). 1–6.

[116]. Zhao Shiquan, Wu Jian, Sheng Victor S., Ye Chen, Zhao Pengpeng, and Cui Zhiming. 2015. Weak labeled multi-label active learning for image classification. In Proceedings of theACM International Conference on Multimedia (ACM MM'15). 1127–1130.

[117]. Zheng Yuhui, Byeungwoo Jeon, Xu Danhua, Wu Q. M. Jonathan, and Hui Zhang. 2015. Image segmentation by generalized hierarchical fuzzy C-means algorithm. J. Intell. Fuzzy Syst28, 2 (2015), 961–973.

[118]. Zhou Zhi Hua. 2010. Semi-supervised learning by disagreement. Knowl. Inf. Syst24, 3 (2010), 415–439.

[119]. Zhou Zhi Hua, Sun Yu Yin, and Li Yu Feng. 2009. Multi-instance learning by treating instances as non-I.I.D. samples. In Proceedings of theAnnual International Conference on Machine Learning (ICML'09). 1249–1256.

[120]. Zhou Zhi Hua and Zhang Min Ling. 2007. Multi-instance multi-label learning with application to scene classification. In Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'07). 1609–1616.

[121]. Zhou Zhi Hua, Zhang Min Ling, Huang Sheng Jun, and Li Yu Feng. 2012. Multi-instance multi-label learning. Artif. Intell176, 1 (2012), 2291–2320.

[122]. Zhu Xiaojin. 2005. Semi-Supervised Learning Literature Survey. Computer Science Technical Report1530, University of Wisconsin-Madison.

CCS Concepts:

- Theory of computation → Active learning;

Multi-Label
Images

Fig. 1.

Several examples of multi-label images: Each image can be labeled with several of predefined labels, such as *beach*, *mountain*, *fall foliage*, *urban*, and *field*.

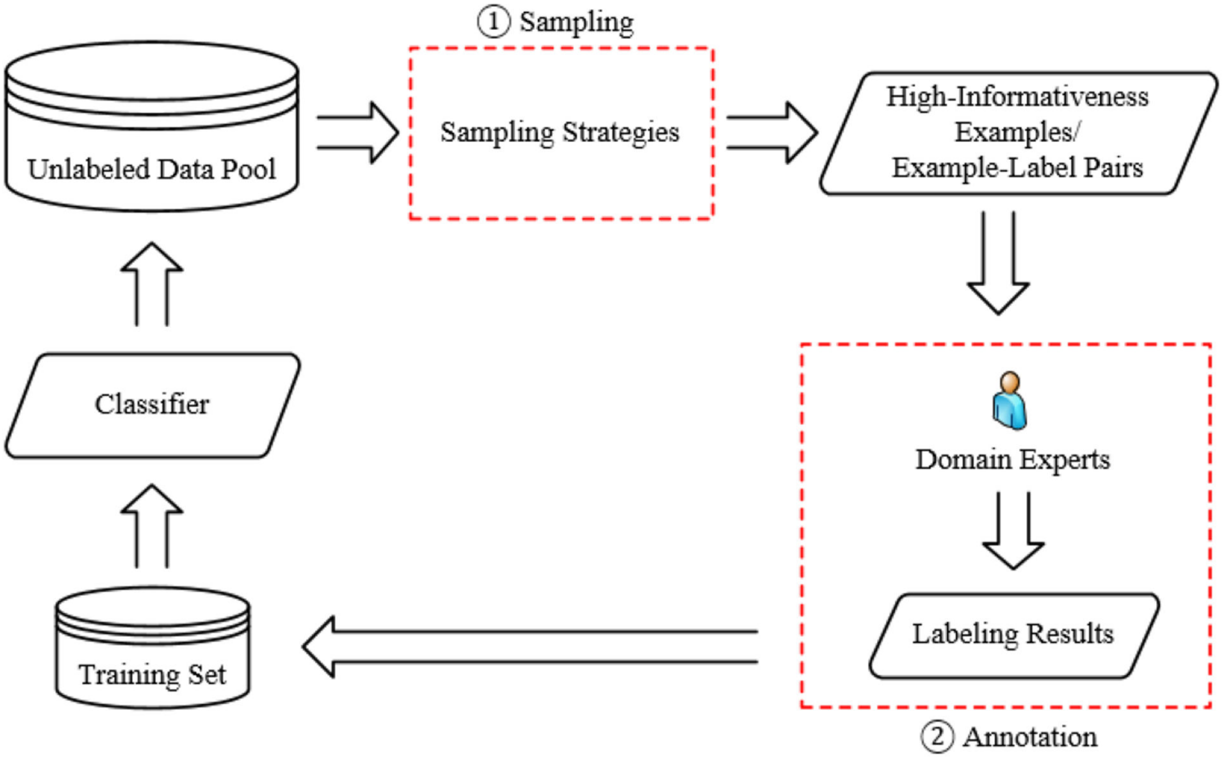


Fig. 2.

A general framework of a typical pool-based multi-label active learning system.

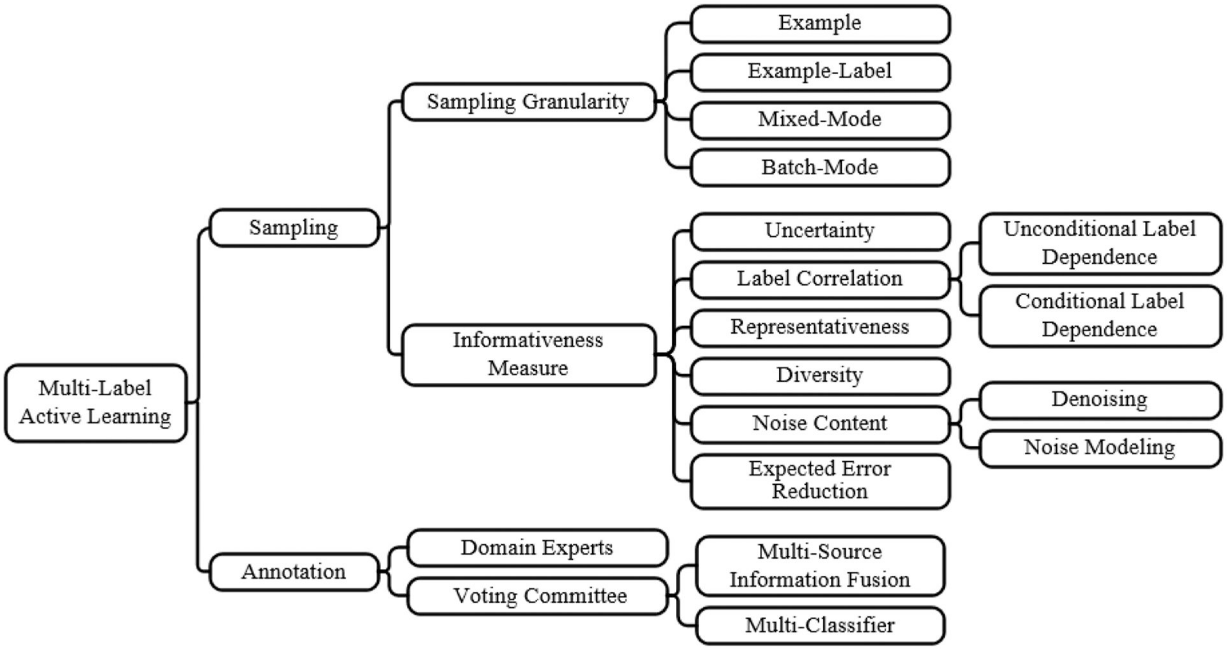
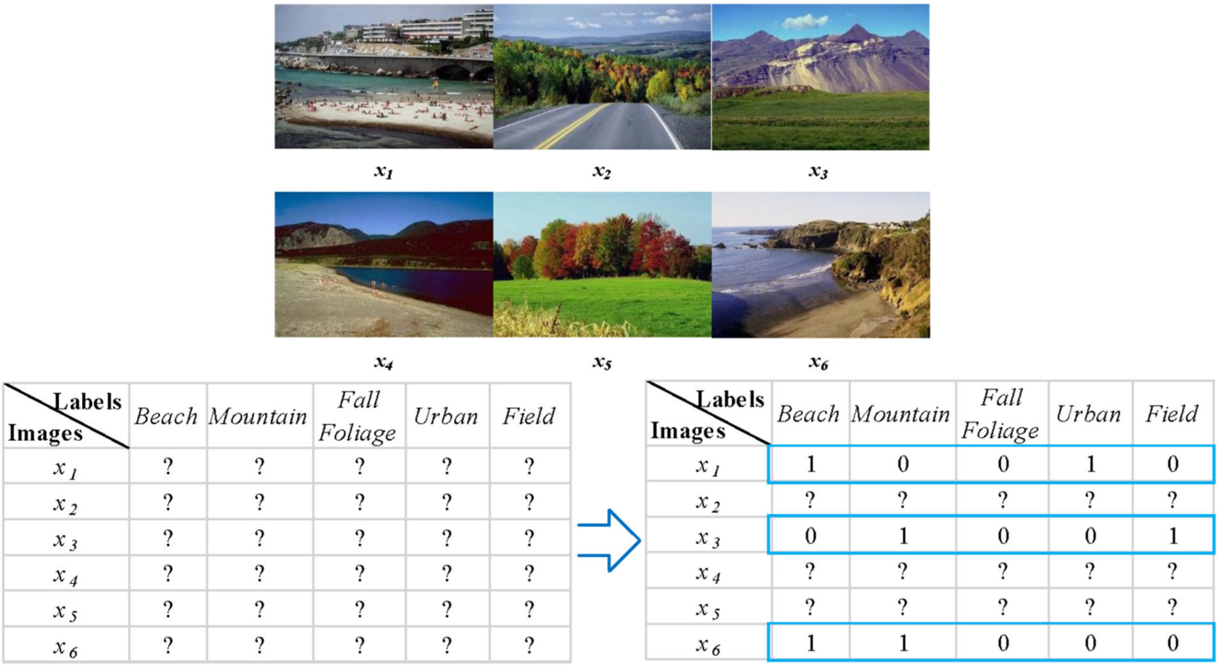


Fig. 3.

A hierarchical structure of existing studies on multi-label active learning. This section focuses on two main research focus areas: sampling and annotation.

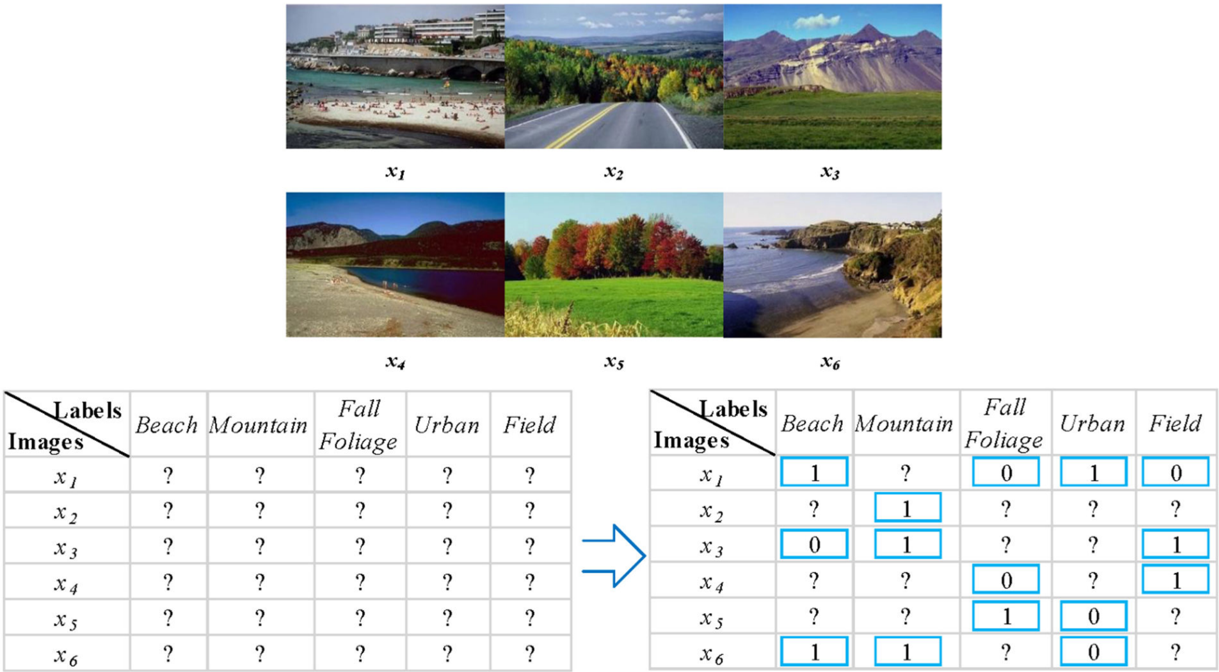


Labels / Images	*Beach*	*Mountain*	*Fall Foliage*	*Urban*	*Field*
x_1	?	?	?	?	?
x_2	?	?	?	?	?
x_3	?	?	?	?	?
x_4	?	?	?	?	?
x_5	?	?	?	?	?
x_6	?	?	?	?	?

Labels / Images	*Beach*	*Mountain*	*Fall Foliage*	*Urban*	*Field*
x_1	1	0	0	1	0
x_2	?	?	?	?	?
x_3	0	1	0	0	1
x_4	?	?	?	?	?
x_5	?	?	?	?	?
x_6	1	1	0	0	0

Fig. 4.

An example: one iteration of an example-based multi-label active leaning method that acquires all the labels for each candidate example.



Labels / Images	Beach	Mountain	Fall Foliage	Urban	Field
x_1	?	?	?	?	?
x_2	?	?	?	?	?
x_3	?	?	?	?	?
x_4	?	?	?	?	?
x_5	?	?	?	?	?
x_6	?	?	?	?	?

Labels / Images	Beach	Mountain	Fall Foliage	Urban	Field
x_1	1	?	0	1	0
x_2	?	1	?	?	?
x_3	0	1	?	?	1
x_4	?	?	0	?	1
x_5	?	?	1	0	?
x_6	1	1	?	0	?

Fig. 5.

An example: One iteration of an example-label-based multi-label active learning method that acquires the label for each candidate example-label pair.

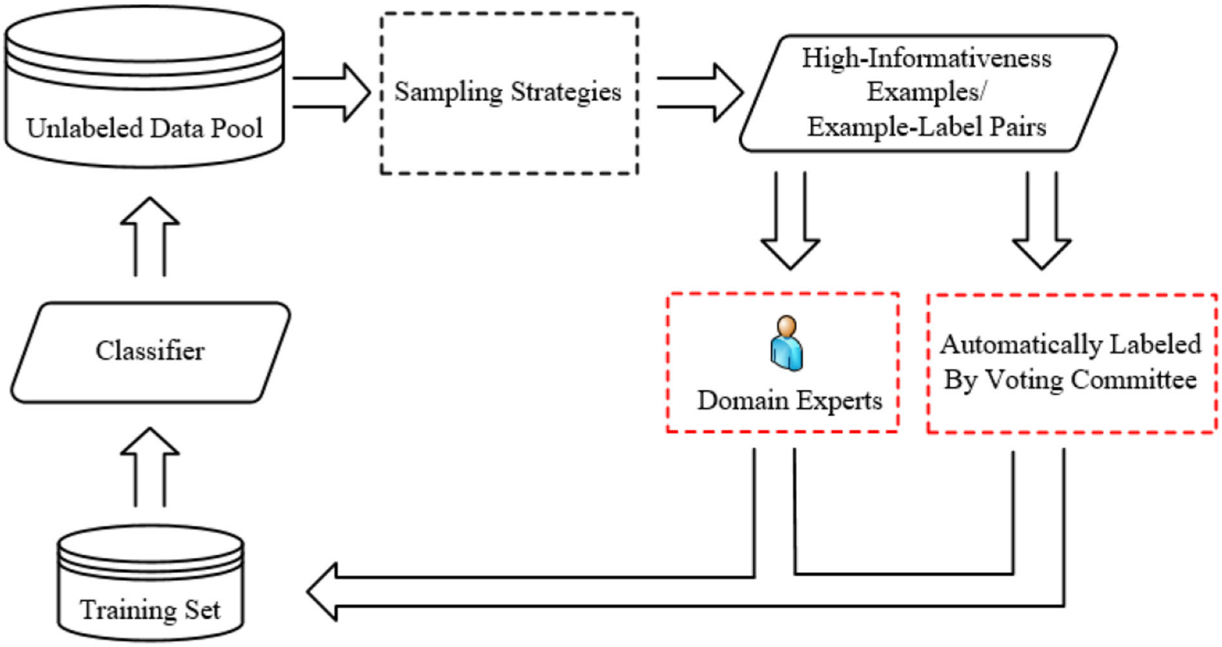

Unlabeled Data Pool
Sampling Strategies
High-Informativeness Examples/ Example-Label Pairs
Classifier
Domain Experts
Automatically Labeled By Voting Committee
Training Set

Fig. 6.

An illustration of how the manual and automatic annotations are integrated into active learning pipeline.

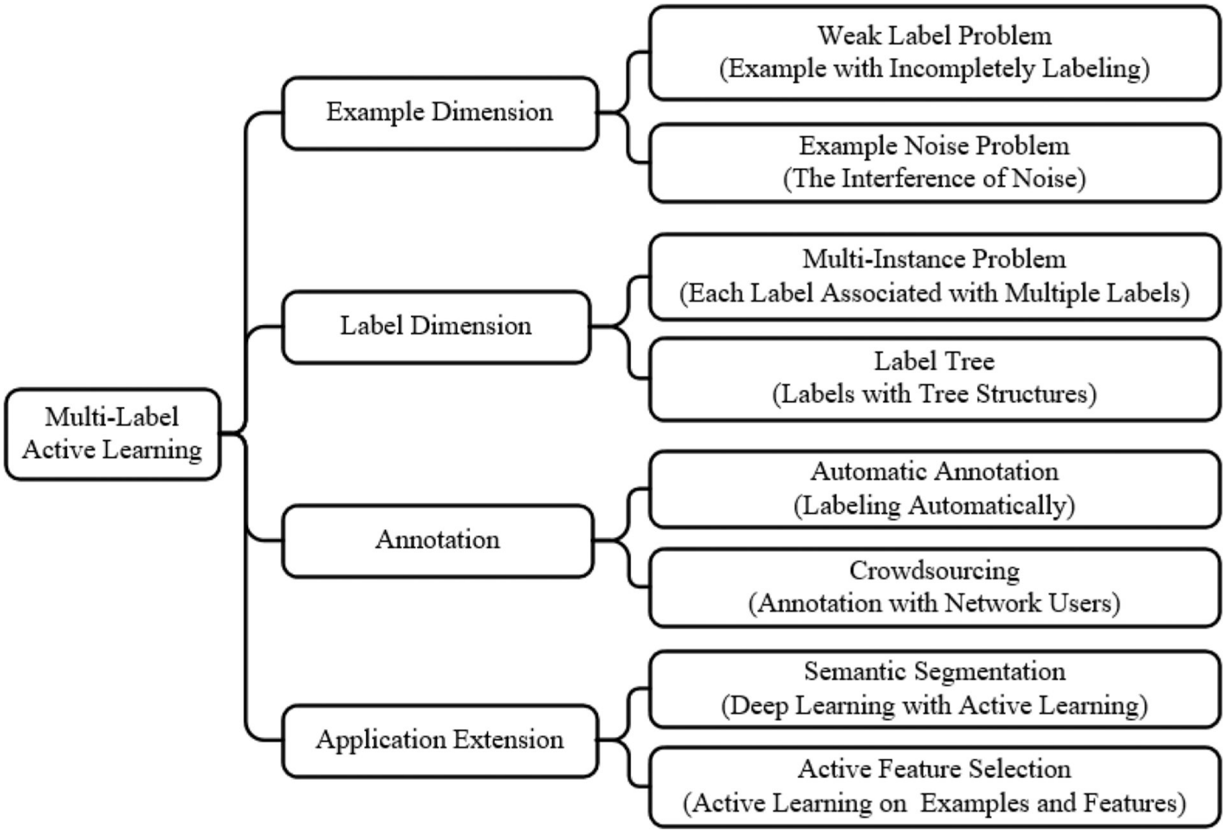

Multi-Label Active Learning
Example Dimension
Weak Label Problem (Example with Incompletely Labeling)
Example Noise Problem (The Interference of Noise)
Label Dimension
Multi-Instance Problem (Each Label Associated with Multiple Labels)
Label Tree (Labels with Tree Structures)
Annotation
Automatic Annotation (Labeling Automatically)
Crowdsourcing (Annotation with Network Users)
Application Extension
Semantic Segmentation (Deep Learning with Active Learning)
Active Feature Selection (Active Learning on Examples and Features)

Fig. 7.

A hierarchical structure of future promises on multi-label active learning in terms of example dimension, label dimension, annotation, and application extension.

Table 1.

Summary of Major Mathematical Notations throughout This Work along with Their Explanations

Notations	Explanations
A	label space with n_m possible class labels $\{l_1, l_2, \dots, l_{n_m}\}$
X	d-dimensional example set
L	a labeled training set
U	an unlabeled data pool
x_i	the ith example with d-dimensional features $\left(v_1^i, v_2^i, \dots, v_d^i\right)$ $(x_i \in X)$
Y_i	the label set associated with the example x_i
$y_{j,k}$	the kth label of example x_j
n_l	the number of examples in L
n_u	the number of examples in U
n_s	the number of selected examples/example-label pairs
Θ	the trained multi-label classifier $\Theta : X \rightarrow A$
$UL(x_j)$	the unlabeled label set of example x_j
$\|UL(x_j)\|$	the number of unlabeled labels in $UL(x_j)$
$LD(x_j)$	the labeled label set of example x_j
$\|LD(x_j)\|$	the number of labeled labels in $L(x_j)$
$c(y_{j,k})$	the correlated label set of $y_{j,k}$ in x_j
$f_k(\cdot)$	the classifier of the kth label l_k in A

Table 2.

Summary of the State-of-the-art Multi-label Active Learning Methods

MLAL Methods	Sampling Granularity				Informativeness Measure						Annotation	
	E	EL	MM	BM	U	LC	R	D	NC	EER	DE	VC
Li et al. 2004 [54]	×									×	×	
Turet al. 2005 [86]	×				×						×	×
Singhet al. 2009 [73]	×				×						×	
Esuli et al. 2009 [20]	×				×						×	
Yang et al. 2009 [102]	×									×	×	
Huang et al. 2010 [34]	×				×		×				×	
Nasierding et al. 2010 [60]	×									×	×	
Chakraborty et al. 2011 [10]	×			×	×			×			×	
Tang et al. 2012 [78]	×									×	×	
Li et al. 2013 [52]	×				×						×	
Jiao et al. 2014 [39]	×			×	×		×	×			×	
Jiao et al. 2015 [38]	×			×	×		×	×			×	
Reyes et al. 2018 [66]	×				×						×	
Reyes et al. 2018 [68]	×			×	×		×	×			×	
Qi et al. 2008 [63]		×			×	×					×	
Qi et al. 2009 [64]		×			×	×					×	
Hua et al. 2008 [33]		×			×	×					×	
Zhang et al. 2009 [114]		×			×	×					×	
Zhang et al. 2010 [115]		×			×	×					×	
Zhang et al. 2012 [112]		×		×	×	×		×			×	
Huang et al. 2014 [35]		×			×	×	×				×	
Zhang et al. 2014 [111]		×		×	×	×		×			×	
Wu et al. 2014 [94]		×			×						×	
Wu et al. 2015 [95]		×			×	×					×	×
Ye et al. 2015 [105]		×			×	×					×	
Yeet al. 2015 [106]		×			×	×					×	
Zhao et al. 2015 [116]		×			×	×					×	
Wu et al. 2016 [98]		×			×	×					×	
Gao et al. 2016 [24]		×		×			×	×			×	
Guo et al. 2017 [30]		×			×	×					×	×
Wu et al. 2017 [91]		×		×	×	×		×	×		×	
Wu et al. 2017 [96]		×			×	×					×	×
Wu et al. 2017 [97]		×			×	×					×	×
Wu et al. 2018 [93]		×			×	×			×		×	
Jiao et al. 2014 [40]			×		×	×					×	
Huang et al. 2015 [36]			×		×	×					×	

The multi-label active learning (MLAL) methods are ordered by the year of their corresponding publication grouped by example-based, example-label-based and mixed-mode-based. **E:** Example-based, **EL:** Example-label-based, **Mmml:** Mixed-mode-based, **Bmml:** Batch-mode-based, **U:**

Uncertainty Metric, **LC:** Label Correlation Metric, **R:** Representativeness Metric, **D:** Diversity Metric, **NC:** Noise Content Metric, **EER:** Expected Error Reduction Metric, **DE:** Domain Experts, and **VC:** Voting Committee.