Chapter 10
Multiple Instance Multiple Label Learning

Abstract As applications grow more complex, proper data representation becomes more relevant. Experience shows that a representation accurately reflecting existing relations and interactions in the data renders the learning task easier to solve. In this context, multiple instance multiple label learning (MIMLL) appears as a flexible learning framework. The combination of MIL and multi-label learning introduces a greater flexibility and ambiguity in the object representation by providing a natural formulation for representing complicated objects. This chapter provides a general introduction to MIMLL. First, a description and formal definition are presented in Sects. 10.1 and 10.2. The main applications are listed in Sect. 10.3. Appropriate evaluation metrics for MIMLL are described in Sect. 10.4. Section 10.5 presents an overview of the proposed methods and Sect. 10.7 describes some current advances. Finally, Sect. 10.6 describes the Yelp classification challenge.

10.1 Introduction

As described throughout this book, MIL is an alternative to traditional single-instance learning and represents a complicated object by a set of instances. Even though it allows to easily describe a complex concept, each observation is assumed to belong to only one class. However, there exist classification scenarios in which samples can belong to several classes. In such a situation, more flexibility needs to be introduced in the representation. In the framework of multiple label learning (MLL) [5], each observation can belong to several classes. Examples include images that belong to several classes simultaneously and text documents classified to several news categories.

In this chapter, MIMLL is described, combining the multi-instance and multi-label perspectives. It is a learning framework that introduces flexibility and ambiguity in the object representation of both the input and output spaces. An object is represented by a bag of instances and is allowed to have multiple class labels. MIMLL combines the MIL and MLL frameworks to formalize objects in real-world problems. For instance, in image classification, an image generally contains several naturally partitioned patches (instances) and the complete image can correspond to multiple semantic

F. Herrera et al., *Multiple Instance Learning*, DOI 10.1007/978-3-319-47759-6_10

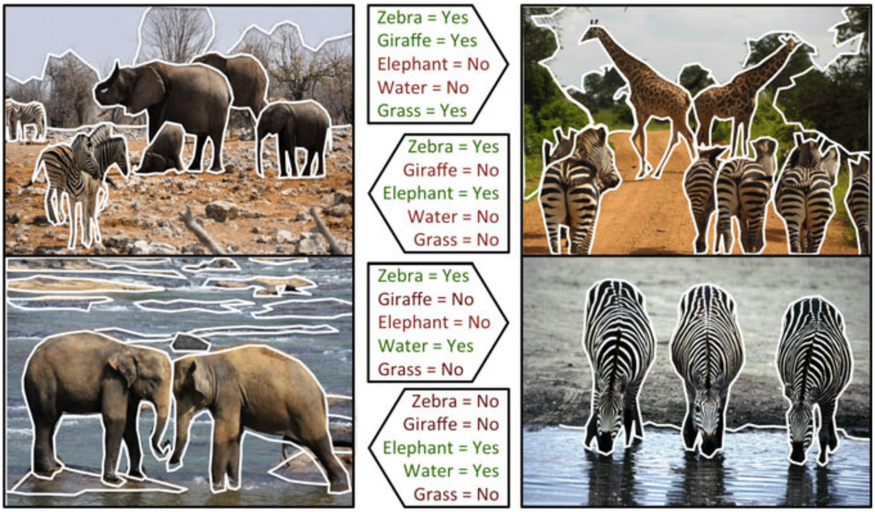


Fig. 10.1 Example of MIMLL problem

classes, such as clouds, grassland, and lions. In bioinformatics, a gene sequence generally encodes a number of segments (instances) and it may be associated with several functional classes, such as metabolism, transcription, and protein synthesis. In text categorization, each document usually consists of several sections or paragraphs (instances), while the document may be assigned to a set of predefined topics, such as sports and Olympic games. In Sect. 10.3, different application domains are described in more depth.

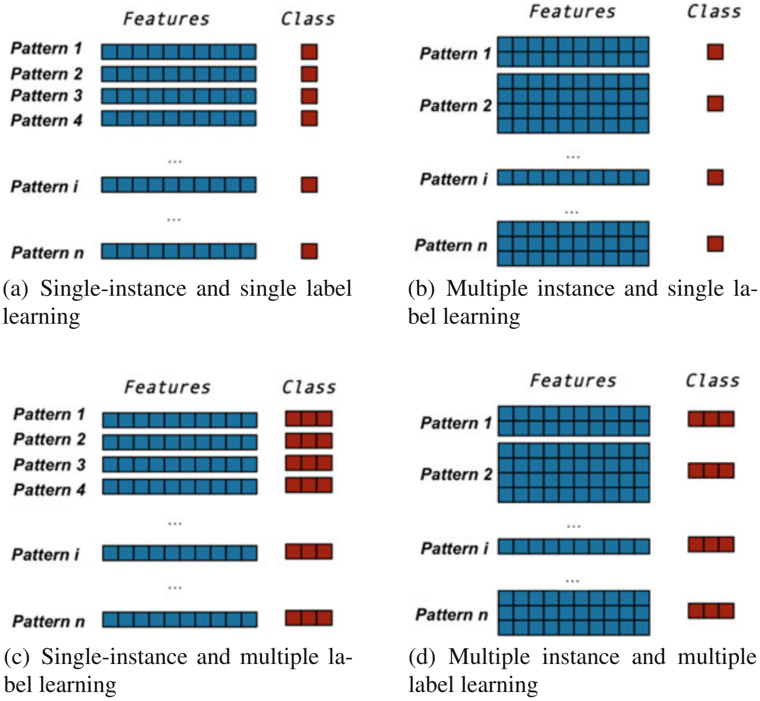
Compared to traditional learning frameworks, MIMLL is more convenient and natural for representing complex objects, because it adds a higher flexibility both in the input space and output space. Figure 10.1 shows an application of image annotation from the MIMLL perspective. Each image is composed of a bag of regions and is associated with multiple labels. The relationship between the image regions and labels is unknown. Concretely, the figure shows four different images where different concepts are considered, such as giraffe, elephant, zebra, water, and grassland. The combination of a multi-label object with a set of instances allows to obtain the relation between the input patterns and their semantic meaning more easily. In some cases, understanding why a particular object has a certain class label is even more important than simply making an accurate prediction. Under the MIMLL representation, we may discover that one object has $label_1$ because it contains $instance_1$ and another has $label_2$ because it contains $instance_2$, while the occurrence of both $instance_1$ and $instance_2$ triggers a more complex concept, such as a particular African region depending on the represented animals and landscape. In this context, MIMLL has demonstrated better performance to discover high-level concepts. For example, the concept of an African zone has a broad connotation and the images belonging to the Africa concept are varied and therefore not easy to classify. However, if we can exploit some low-level sub-concepts that are less ambiguous and easier to learn, such as water, grass, elephant, zebra, and giraffe, it is possible to induce the portrayed area of Africa much easier than by learning it directly.

10.2 Formal Definition

As a preliminary step to define MIMLL, we study its relationship with single-instance learning, multi-instance learning and multi-label learning, focusing on classification. The definitions of single-instance learning and MIL can be consulted in Chaps. 1 and 2. Figure 10.2 shows the differences among the different learning frameworks.

In single-instance learning, an instance x is a point in the instance space $\mathbb{X}$. It is commonly assumed that $\mathbb{X} \subseteq \mathbb{R}^d$, that is, each instance is described by a vector of d elements. The space $\mathbb{X}$ can be generalized to $\mathbb{X} \subseteq \mathscr{A}^d = \mathscr{A}_1 \times \cdots \times \mathscr{A}_d$ so that each instance is described by a d-dimensional vector where each attribute $\mathscr{A}_i (i = 1, \ldots, d)$ takes values from a finite or infinite set $\mathscr{V}_i$.

In MIC, a bag X is a set of n instances $\{x_1, \ldots, x_n\}$, $x_i \in \mathbb{X}$, $\forall i \in [1, \ldots n]$. Each bag can contain a distinct number of instances. In a training set $D = (\mathbf{X}, \mathbf{L})$,



(a) Single-instance and single label learning

(b) Multiple instance and single label learning

(c) Single-instance and multiple label learning

(d) Multiple instance and multiple label learning

Fig. 10.2 Learning frameworks

$\mathbf{X} = \langle X_1, \ldots, X_m \rangle$ is a set of m bags and $\mathbf{L} = \langle \ell_1, \ldots, \ell_m \rangle$ is a set of class labels. Each bag X_i is assigned a class label $\ell_i \in \mathbb{L}$ for all $i = 1, \ldots, m$. The classes of instances inside the bags are not known. The objective is to find a function $f_{MIC} : \mathbb{N}^{\mathbb{X}} \to L \in \mathbb{L}$, that allows us to predict class labels of new bags as accurately as possible. This problem can be seen as multi-instance single label learning.

On the other hand, MLL describes each object by one instance associated with several class labels. In a classification problem, we have a training set $D = (\mathbf{X}, \mathbf{L})$, where $\mathbf{X} = \langle x_1, \ldots, x_m \rangle$ is a set of m instances and $\mathbf{L} = \langle L_1, \ldots, L_m \rangle$ is a set of m class label sets. Each instance x_i is assigned a set of class labels $L_i = \langle \ell_{i1}, \ldots, \ell_{ik_i} \rangle$, with $\ell_{ij} \in \mathbb{L}, \forall j \in [1, \ldots, k_i]$. The objective is to find a function $f_{MLL} : \mathbb{X} \to L \subseteq \mathbb{L}$ that assigns a combination of labels to each instance. This problem can be seen as single-instance multi-label learning.

Based on the previous definitions and according to the formulation given by Zhou et al. [39], the task of MIMLL would consist of learning a function $f_{MIML} : \mathbb{N}^{\mathbb{X}} \to L \subseteq \mathbb{L}$ from a set of MIML training examples $\{(X_i, L_i) | 1 \leq i \leq m\}$, where $X_i \subseteq \mathrm{X}$ is a bag of n_i instances $X_i = \langle x_{i1}, x_{i2}, \ldots, x_{in_i} \rangle$ and $L_i \subseteq \mathrm{L}$ is a set of k_i labels $L_i = \langle \ell_{i1}, \ell_{i2}, \ldots, \ell_{ik_i} \rangle$ associated with X_i.

10.3 Applications

There are many real-world problems which can be properly formalized under MIMLL, since their complex objects involve a representation ambiguity in the input space (an object can have many input descriptions) and output space (an object can belong to many classes). Most of them are based on applications studied in Sect. 2.4, although each object is now represented not only by a set of instances but also by a set of labels.

10.3.1 Image Classification

Image classification is one of the most widely studied MIMLL applications. The purpose is to, given a image, identify the objects or categories that are portrayed. Traditional studies have used global image features to solve this task. Such features cannot characterize an image well, since it is usually composed of several complex objects. MIMLL represents an image as a bag of instances, where each instance corresponds to an image region. These image regional features can better characterize complex contents. On the other hand, assigning a single label to an image may be impractical in real applications. MIMLL achieves a more appropriate representation by associating multiple labels with an image. The learning aim is to uncover the unknown relationship between the regions and class labels. The learned relationship can be used to classify unlabeled images.

Region-based image classification of natural scene images has been addressed in several works [1, 4, 14, 33–37]. These studies employ 2000 images and five categories (desert, mountains, sea, sunset, and trees). This task has become a benchmark for image annotation. The classification objective is to predict which categories the complete image represents. Only 22 % of the images in the dataset belong to more than one class. The average number of labels per image is 1.24 ± 0.44. Each image is represented as a bag of nine 15-dimensional instances (image patches).

Other works like [19, 22] use the classic Corel dataset containing 5000 images. The whole set consists of 50 groups, such as beach, aircraft, and tiger. Each group contains 100 similar images and every image is annotated with one to five categories. The total number of keywords in the Corel dataset is 371.

10.3.2 Video and Audio Concept Detection

With the rapid development of storage devices, networks and compression techniques, large collections of digital videos are available. Automatic video annotation has emerged as an interesting topic in the multimedia research community to facilitate the annotation of videos with concepts describing the information in the video content at the semantic level. These concepts can be used to index or browse the video.

Traditional studies represent one video clip with a flat feature vector. However, video data usually has a natural hierarchical structure. A video can be represented by a hierarchy including, from large to small: shot, frame, and region within the frame. Moreover, a video clip is generally relevant to multiple concepts. MIMLL represents each shot as a bag of instances in which each instance corresponds to a key-frame of the video. The relation between instances plays an important role, for example when the number of key-frames containing the concept needs to be determined in order to predict whether the shot is associated with that particular concept.

Xu et al. [29] work with 170 h of TV news videos from 13 different programs in English, Arabic, and Chinese to detect the presence or absence of 10 predetermined benchmark concepts in each shot. These concepts are walking, running, explosion fire, maps, flag US, building, waterscape waterfront, mountain, prisoner, sports, and car.

The automatic recognition of bird species from audio files has been dealt with in MIMLL as well. Habitat loss, declining biodiversity, and climate change require the development of better tools to monitor birds, including their ranges, diversity, and phenology. Birds are a good indicator of ecosystem health and diversity, because they are relatively easy to detect, may provide information about other organisms (plants, insects…) and respond quickly to environmental change. However, monitoring bird populations and activity is an intensive task. Machine learning tools can be used instead to estimate species presence/absence, abundance, gender, age, and other individual characteristics. In MIMLL, each audio record is a bag of instances and each instance is a segment of the spectrogram corresponding to syllables of bird sounds described by a feature vector of acoustic properties. The labels are the species present in the recording.

Briggs et al. [3] and Pham et al. [17] work with more than 10 terabytes of audio recordings of birds using unattended omnidirectional microphones. These microphones pick up all sounds in the environment, particularly wind and stream noise. There are often several birds vocalizing at once. The goal is to detect the presence or absence of 13 different species of birds. Each recording contains between one and five species, with 2.144 species on average.

10.3.3 Text Categorization

Another application domain of MIMLL is text categorization. Traditional studies represent a whole document by means of a word bag. However, a document usually consists of several separated semantic parts (paragraphs). Different topics evolve along these parts. MIMLL represents each document as a bag of instances, where each instance corresponds to a paragraph in the document or a text segment enclosed in a sliding window of a particular size. Different labels are assigned to each document.

Several works deal with fragment-based text classification [1, 14, 31, 34–36]. Although all of them are based on the classic Reuters-21578 text collection, a benchmark for text categorization, different configurations have been used to represent documents in the MIMLL framework. The original dataset contains 10788 and 10 classes, but the most commonly used dataset contains 2000 documents and the aim is to categorize them in seven different categories. Documents with multiple labels comprise around 15 % of the dataset and the average number of labels per document is 1.15 ± 0.37.

10.3.4 Bioinformatics

Common bioinformatics tasks are the understanding of gene functions, interactions and networks. Nature often brings several domains together to form multi-domain and multi-functional proteins. Each domain may fulfill its own function independently or together with its neighbors. With the rapid growth of the number of sequenced genomes, the vast majority of proteins can only be annotated computationally. A gene sequence generally encodes a number of segments, each one of which can be expressed as an instance in MIMLL. The gene sequence itself may be associated with several functional classes, such as metabolism, transcription, and protein synthesis.

Several works carry out the automated annotation of protein functions [26, 28]. They use a complete proteome on seven real-world organisms, containing 379 proteins (bags) with a total of 320 gene ontology terms (classes) given by the Gene Ontology Consortium. From the MIMLL perspective, each protein is represented as a bag of instances, where each instance corresponds to a domain and is labeled with a group of gene ontology molecular function terms. The average number of instances (domains) per bag (protein) is 3.20 ± 1.21 and the average number of labels per example (protein) is 3.14 ± 3.33.

Li et al. [13] carry out the automated annotation of embryo images (concretely, studies of Drosophila embryogenesis). They use six different ranges to classify the gene expressions captured in the images with anatomical and development ontology terms. Each image contains only one individual embryo represented by a bag. The image is divided in several patches using a 128-dimensional vector to represent each patch.

10.4 Evaluation Metrics

MIMLL algorithms make multi-label predictions. Their performance is evaluated with multi-label metrics that also have to consider that the dataset consists of bags of instances.

Similar to MLL [5], *example-based metrics* are calculated separately for each bag and averaged over samples, while the *label-based metrics* are computed independently for each label before averaging. Two different strategies can be applied, namely *macro-averaging* and *micro-averaging*. In the former, the metric is calculated individually for each label and the result is divided by the number of labels. For the latter, the hit and miss counts for each label are first aggregated and the metric is computed only once after that. The metrics can also be grouped according to the result provided by classifier. In *binary bipartition*, a vector of 0s and 1s, indicating which of the labels are relevant to each sample, is obtained. In *label ranking*, a label list ranked according to some relevance measure is returned.

In this section, we describe five popular measures. These are example-based metrics to evaluate bipartitions. With respect to notation, D is a MIML dataset, $D = (X, L)$, where X is a set of n bags $X = \{X_1, \ldots, X_n\}$. Each bag $X_i = \{x_{i1}, \ldots, x_{in_i}\}$ is composed of n_i instances and $L = \{L_1, \ldots, L_n\}$ is a set of n label sets, where each label set $L_i = \{\ell_{i1}, \ldots, \ell_{ik}\}$ is composed of k labels. The function $h(X_i)$ returns a set of labels of X_i. The $|\cdot|$ operator counts the number of elements in a set.

- **Hamming loss**: this metric counts the number of incorrect example-label pairs,

$$Hloss = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}|h(X_i)\triangle L_i|,$$

  where $\triangle$ denotes the symmetric difference between the two sets L_i, the real label set of the ith bag, and $h(X_i)$, the predicted one. There are l labels in total. The Hamming loss, which should be minimized, is an indicator of the errors of the classifier proportional to the label set length. It results in different assessments for the same amount of errors depending on the label set lengths of the dataset.

- **Accuracy**: the ratio between the number of correctly predicted labels and the total number of active labels, both in the real label set and the predicted one, is evaluated. Like all example-based metrics, the accuracy is computed for each instance and then averaged, namely

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{|L_i \cap h(X_i)|}{|L_i \cup h(X_i)|}.$$

- **Precision**: this measure computes the ratio of the number of correctly predicted labels and the total number of predicted labels. It can be interpreted as the percentage of predicted labels that are truly relevant for the bag. It is calculated as

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|L_i \cap h(X_i)|}{|h(X_i)|}.$$

- **Recall**: the ratio of the number of correctly predicted labels and the total number of real labels is evaluated. Recall can be interpreted as the percentage of correctly predicted labels among all truly relevant labels, that is,

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{|L_i \cap h(X_i)|}{|L_i|}.$$

- **F1 score**: this metric, also known as the F-measure, is based on the precision and recall statistics. The mean F1 score is obtained by averaging the F1 scores of the individual labels. It is a weighted measure of how many relevant labels are predicted and how many of the predicted labels are relevant. It is computed as

$$F1Score = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \cdot |h(X_i) \cap L_i|}{|h(X_i)| \cap |L_i|}.$$

10.5 Multi-instance Multi-label Learning Methods

MIMLL methods are classified according to the general grouping proposed by Zhou et al. [39]. A distinction is made between algorithms that solve the problem by degeneration or those that solve it by regularization. In degeneration methods, the problem is transformed to a MIL or MLL task. In regularization algorithms on the other hand, the problem is addressed directly using the MIML representation.

10.5.1 Methods Based on Problem Degeneration

These methods use an intuitive way to tackle the problem by identifying its equivalent in traditional supervised learning (that is, single-instance and single label learning, SISL) via problem reduction. Both MIL and MLL are degenerate versions of MIMLL. They are used as a bridge to solve the MIML problem. Based on this idea, two different paradigms have been proposed.

- **MIL as a bridge**: these models transform the MIMLL task, which learns a function $f_{MIML} : \mathbb{N}^{\mathbb{X}} \rightarrow 2^{\mathbb{L}}$, to a MIC task learning a function $f_{MIC} : \mathbb{N}^{\mathbb{X}} x \; \mathbb{L} \rightarrow \{-1, +1\}$. For any $\ell \in \mathrm{L_i}, f_{MIC}(X_i, \ell) = +1$ if $\ell \in L_i$ and -1 otherwise. The labels L^* for a new example X^* can be determined as $L^* = \{\ell \mid sign[f_{MIC}(X^*, \ell)] = +1\}$. As an illustration, Fig. 10.3 shows the transformation of a MIML problem with three labels into three different MIC problems with one label each. The resulting MIC task could be transformed into a traditional supervised learning task to learn a function $f_{SISL} : \mathbb{X} \rightarrow \mathbb{L} \in \{-1, +1\}$, under a constraint specifying how to derive $f_{MIC}(X_i, \ell)$ from $f_{SISL}(x_{ij}, \ell)(j = 1, \ldots, n_i)$. For any $\ell \in \mathrm{L_i}, f_{SISL}(x_{ij}, \ell) = +1$ if $\ell \in L_i$ and -1 otherwise. The constraint can be $f_{MIC}(X_i, \ell) = sign\left[\sum_{j=1}^{n_i} f_{SISL}(x_{ij}, \ell)\right]$, which is used to transform MIC tasks into traditional supervised learning tasks.

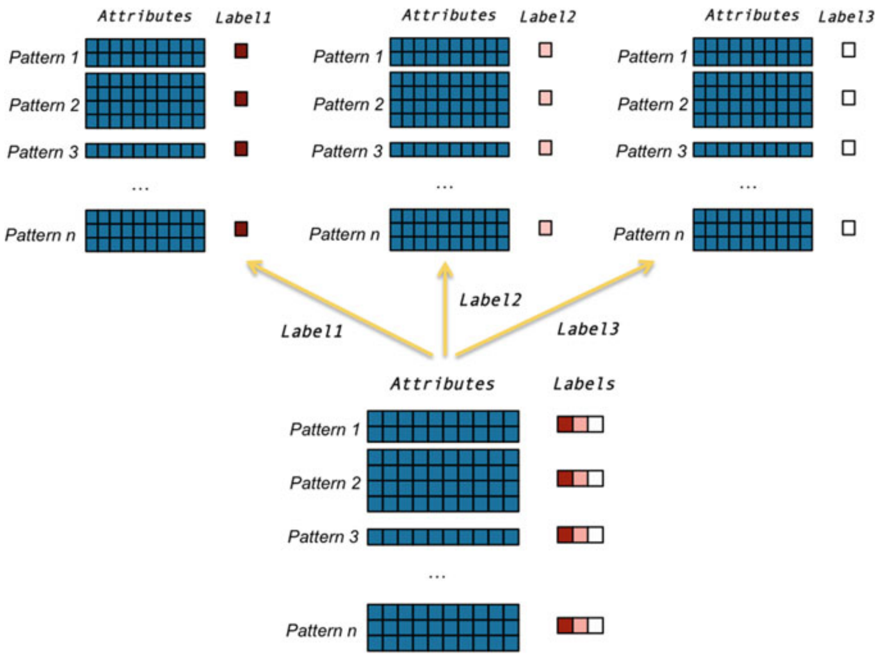
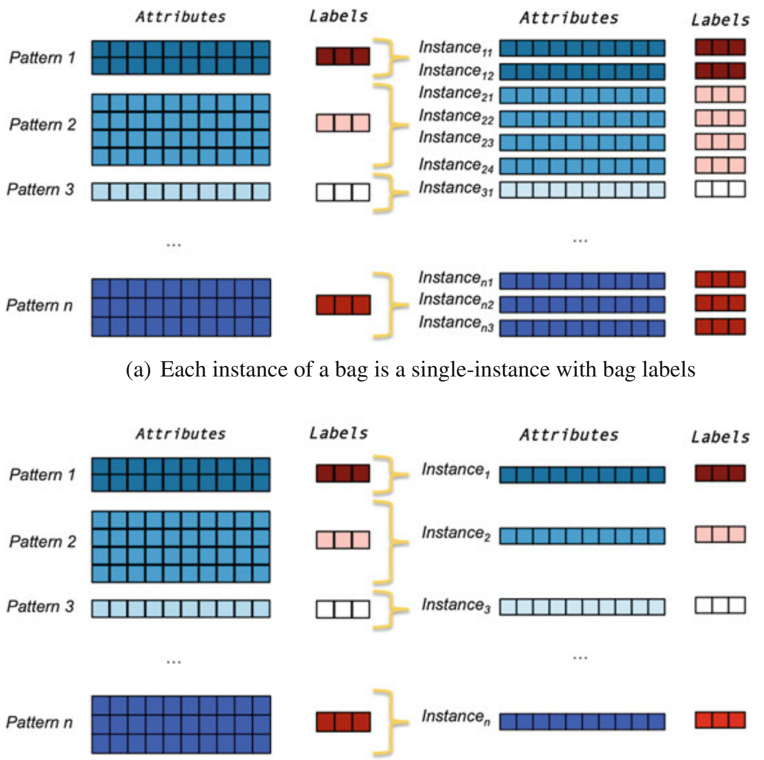


Fig. 10.3 Using MIL as bridge to solve MIMLL problem

- **MLL as a bridge**: these methods transform a MIMLL task into a MLL task, that learns a function $f_{MLL} : \mathbb{Z} \rightarrow 2^{\mathbb{L}}$. For any $z_i \in \mathbb{Z}$, $f_{MLL}(z_i) = f_{MIML}(X_i)$ if $z_i = \phi(X_i)$, $\phi : \mathbb{N}^{\mathbb{X}} \rightarrow \mathbb{Z}$. The labels for a new example X^* can be determined as $L^* = f_{MLL}(\phi(X^*))$. The mapping ϕ can be any that encodes bags as single vectors. As an example, Fig. 10.4 shows two possible transformations. In Fig. 10.3a, each instance in a bag is converted into an instance with the same labels, while Fig. 10.3b depicts the situation where each bag is converted to one instance using as function ϕ returning the closest instance to the bag centroid. In the latter case, each bag yields one pattern. The MLL task can be transformed into a traditional supervised learning task learning a function $f_{SISL} : \mathbb{Z} x \mathbb{L} \rightarrow \{-1, +1\}$. For any $\ell \in \mathrm{L_i}$, $f_{SISL}(z_i, \ell) = +1$ if $\ell \in L_i$ and -1 otherwise, such that $f_{MLL}(z_i) = \{\ell \,|\, f_{SISL}(z_i, \ell) = +1\}$.

Table 10.1 shows an overview of algorithms developed within this scheme. A distinction between them is made based on the degeneration scheme and on the algorithm type used to solve the problem.



(a) Each instance of a bag is a single-instance with bag labels

(b) The closest instance to bag centroid is used as single-instance with bag labels

Fig. 10.4 Using MLL as bridge to solve MIMLL problem

Table 10.1 Models based on problem degeneration

Multi-label learning as brigde
Kernel-based methods
MIMLSVM [37]
MIMLSVM$^+$ [13]
E-MIMLSVM$^+$ [13]
SISL-MIML [10]
Ensemble methods
En-MIMLSVM [29]
Neural Networks-based Methods
CPNMIML [30]
Multi-instance learning as brigde
Ensemble methods
MIMLBOOST [37]

The first subgroup consists of *kernel-based methods*. Zhou et al. [37] published one of the pioneering works in this area. They proposed the MIMLSVM method, which solves a MIML problem by degenerating it into a single-instance multi-label problem through a clustering process. Li et al. [13] proposed two different approaches based on SVMs. The first one, MIMLSVM$^+$, employs a degeneration strategy that decomposes the learning of multiple labels into a series of binary classification tasks. An SVM is constructed for each of them. Their second method, E-MIMLSVM$^+$, extends MIMLSVM$^+$ by incorporating the term correlations via kernel-based multi-task learning techniques. An improved degeneration approach is defined by Nguyen et al. [10], where the authors propose an SISL-MIML algorithm based on SVM. They use quadratic and integer programming to solve the problem.

Ensemble methods are also encountered. Zhou et al. [37] were one of the first to propose a solution to the MIML problem by degenerating it into a multi-instance single-label problem. Their MIMLBOOST method reduces the problem by adding pseudo-labels to every instance. Xu et al. [29] proposed the En-MIMLSVM algorithm based on the MIMLSVM method. It is an ensemble that first samples several subsets from the majority class independently. It then trains multiple classifiers using these subsets and the minority class. All constructed classifiers are combined to obtain the final decision. With this methodology, En-MIMLSVM is able to deal with class imbalance.

With respect to *neural networks-based methods*, Yan et al. [30] proposed the CPN-MIML algorithm that combines probabilistic latent semantic analysis (PLSA) with the neural networks. Concretely, the PLSA model translates the MIML problem into a single-instance multi-label problem. A neural network method is used to solve it.

The main shortcoming of degeneration models is that they do not use any information about connections between instances and labels or correlations among labels. This information is lost during the reduction process, although it can help improve the performance of algorithms. On the one hand, compared to the MLL framework, MIMLL could capture the intrinsic causation of each individual label and directly model the latent semantic meaning of instances. On the other hand, in contrast with MIL methods that model individual labels independently, MIMLL can simultaneously model the labels as well as their interactions.

10.5.2 Methods Based on Problem Regularization

As stated above, the performance of degeneration algorithms may suffer from the information loss incurred during the reduction process. Ideally, the connections between instances and labels as well as the correlations among labels should be taken into account. This group of methods includes the remaining regulation frameworks that have been proposed to solve MIMLL problems. Table 10.2 shows an overview.

Table 10.2 Models based on problem regularization

Maximum margin-based methods
M3MIML [34]
MIMLwel [32]
Neural networks-based methods
MIMLRBF [35]
IMIMLRBF [14]
IMIMLRBF-GMBO [1]
MIMLNN [4]
Nearest neighbor-based methods
MIML-kNN [36]
Markov-MIML-kNN [25]
Kernel-based methods
D-MIMLSVM [38]
ML_MLML [24]
Ensemble methods
Peng et al. [19]
EnMIMLNN [26]
Other methologies
Yang et al. [31]
MIML-RE [23]
MIMLGP [6]
Pham et al. (I) [16]
Pham et al. (II) [17]

Following the same procedure as above, these algorithms are grouped according to the approach used to solve the MIMLL problem.

Maximum margin-based methods generally use a subset of the available instances in a given bag and maximize the margin between classes. The score of a bag with respect to each class is computed from the score-maximizing instance in the bag. One of the earliest works in this context was of Zhang et al. [34], who proposed a maximum margin method named Maximum Margin Method for Multi-Instance Multi-Label learning (M3MIML). This method directly considers the connections between instances and labels by defining a specific margin on each example. M3MIML assumes a linear model for each class, where the output for one class is set to the maximum prediction of all the MIML examples instances with respect to the corresponding linear model. Subsequently, the outputs for all possible classes are combined to define the margin of the MIML example within the classification system. Following a similar theory, Yang et al. [32] proposed the MIMLwel approach, that assumes that highly relevant labels share some common instances and that the underlying class means of bags for each label have a large margin. In this proposal, a bag of instances is first mapped to a feature vector, where each element measures the degree of the bag associated with a group of similar instances. Afterward, sparse predictors are employed to learn the bag labels such that the class means of bags for each label are maximized.

Proposals based on *neural network methods* for tackling MIML problems have been developed as well. Zhang et al. [35] proposed the MIMLRBF algorithm, which uses a radial basis function (RBF). A k-medoids clustering step groups the examples of each class. The weights of the method are optimized by a sum-of-squares error function. An improved version of this model was proposed by Li et al. [14]. Their IMIMLRBF method applies an improved k-medoids clustering on the data that still performs appropriately in case of noise. Another improvement of MIMLRBF has recently been developed [1]. The authors proposed a hybrid search method to estimate the RBF neural network parameters (the weights, widths and centers of the hidden units) simultaneously. First, the Gases Brownian Motion optimization algorithm is used to determine the width and center of the network nodes. Next, the parameters are optimized by a gradient-based method. Chen et al. [4] also proposed a multi-instance multi-label algorithm based on neural networks, MIMLNN, based on the popular multi-layer perceptron and derived with the classic backpropagation algorithm.

Proposals based on *k-Nearest Neighbor* are also used to solve this type of problems. Zhang et al. [36] proposed the MIML-kNN algorithm based on the popular k-nearest neighbor technique. MIMLkNN makes predictions based on neighboring and citing examples. This algorithm was computationally optimized with MarkovMIML-kNN learning [25]. MarkovMIMLkNN is a nearest neighbor approach to learn correct labels based on neighbor information as well as on the affinities in a Markov chain. The Markov chain computes the class probability of each object, instead of determining the k-nearest neighbors of the unseen object and using maximum a posteriori probability to calculate its label.

With respect to *methods based on kernels*, Zhou et al. [38] proposed the D-MIMLSVM algorithm using SVMs. Its basic assumption is that the labels associated with the same examples are somehow related and that the bag classification performance depends on the information loss between the labels and the predictions on the bags as well as on the constituent instances. Recently, Tong et al. [24] proposed the ML_MLML algorithm. This method proceeds in three steps. First, instance correlations in a bag are described by constructing a graph. This graph is mapped to a vector in a high-dimensional space to represent the bag features. With this information, the multi-instance bag is transformed into a single-instance sample. Next, considering that predictions of different labels correspond to graphs in different scales, MK_MIML introduces multi-kernel fusion. It constructs multiple kernel functions according to different parameters and graphs in different scales. In the fusion step, a convex combination of the kernels is considered. Finally, the algorithm performs its classification by means of SVM.

Several proposals use *ensemble-based methods*. Peng et al. [19] proposed an ensemble method to combine the results of MIMLSVM$^+$ trained on different visual features. More recently, Wu et al. [26] proposed an ensemble MIML learning framework, EnMIMLNN. Concretely, three algorithms were developed by combining the advantage of three kinds of Hausdorff distance metrics and different voting-based methods.

The remaining frameworks to address the MIML problem are grouped together. Yang et al. [31] proposed the Dirichlet–Bernoulli Alignment (DBA) approach, a probabilistic generative model for multi-class, multi-label, and multi-instance corpora. DBA assumes a tree-structure in the data. Its model is similar to latent Dirichlet Allocation. In DBA, each pattern is modeled as a mixture over the set of predefined classes. An instance is then generated independently conditioned on a sampled class label. The label of a pattern is generated from a Bernoulli distribution conditioned on all the sampled labels used for generating its instances. From another perspective, Surdeanu et al. [23] proposed MIML-RE, a graphical model based on distant supervision for relation extraction. It models both multiple instances (by modeling the latent labels assigned to instances) and multiple labels (by providing a simple method to capture dependencies between labels). The proposal of Briggs et al. [2] presented a possible solution using label ranking. They proposed rank-loss support instance machines, that optimize a regularized rank-loss objective for each bag and can be instantiated with different aggregation models connecting instance-level and bag-level predictions. He et al. [6] proposed the MIMLGP algorithm based on a Gaussian process. The basic idea is to define a latent function with a Gaussian process prior in the instance space for every label and then output the probabilities over different labels for each sample based on the latent function values of its instances. In later work, MIMLGP was used to solve multi-label problems in visual mobile robot navigation [7]. Recently, models based on the maximum likelihood approach have been developed. Pham et al. [16] proposed a discriminative probabilistic model based on maximum likelihood to determine the model parameters and learn an instance-level classifier that accounts for novel instances. At the same time, Pham et al. [17] proposed a graphical model based on these principles taking into account the inner structure of each class.

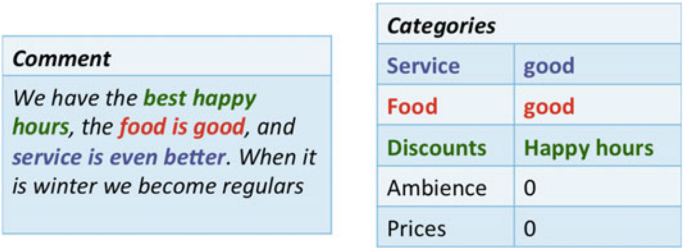
10.6 Case Study: Kaggle Yelp Challenge

The Yelp Restaurant Photo Classification recruitment competition[1] ran on Kaggle from December 2015 to April 2016 corresponding with round 6 of the Yelp dataset challenge. The Yelp Data Challenge is globally organized and consists of the classification of restaurants based on images that various Yelp users have posted. The idea is to use business images to automatically capture meta-data and be able to semantically infer coherent information regarding restaurants, which allows to improve recommendations to users.

Yelp has millions of photos uploaded from all around the world. Some examples are shown in Fig. 10.5. These pictures can provide valuable information and insights into the restaurants they are visually describing. A user may want to know if a restaurant is good for a romantic date, has live music, or serves alcohol. Currently, restaurant labels are manually selected by Yelp users when they submit a photo. They can give ratings and write reviews on businesses and services. While ratings are useful to communicate the overall experience, they do not convey the context which led a

Fig. 10.5 Restaurant photos of the Yelp dataset

[1] https://www.kaggle.com/c/yelp-restaurant-photo-classification.



Comment
We have the best happy hours, the food is good, and service is even better. When it is winter we become regulars

Categories	
Service	good
Food	good
Discounts	Happy hours
Ambience	0
Prices	0

Fig. 10.6 Relating categories with comments

reviewer to that experience. For example, Fig. 10.6 considers a comment about a restaurant given by a Yelp user: *"We have the best happy hours, the food is good, and service is even better. When it is winter we become regulars."* Together with the comment, the user gave the restaurant a 4 star rating. This comment allows to identify that the review talks about *food*, *service* and *deals/discounts* (happy hour). Food and service categories are easy to interpret. Deals and discounts categories correspond to offers during happy hour or specials run by the venue. Other categories, such as an ambiance category related to the look and feel of the restaurant or a price category are not considered in this comment.

This high-level categorization of reviews into relevant categories can help a user to understand the rating assigned by others. It can assist other Yelp users to make a personalized choice, especially when one does not have much time to peruse reviews. It can also be used to rank restaurants according to these categories.

This task can be viewed as a MIMLL problem in the image domain. Each restaurant has an arbitrary number of photos associated with it and can be assigned to multiple categories (many output labels).

10.6.1 Dataset of Round 6 Yelp Challenge

The full dataset is comprised of approximately 234000 images corresponding to 2000 restaurants. The number of images corresponding to each restaurant ranges from 1 to 2974, with roughly 117 images per restaurant on average. The test set contains 237152 images with information on around 10000 businesses. Each business can have nine self-explanatory attributes which are not evenly distributed. The frequency, identifier, and name of each individual label in the training set is presented in Table 10.3.

The goal is to predict class labels from photos uploaded by users. These labels are annotated by the Yelp community and are based on a real-life scrape of Yelp data. Labels can be incomplete or noisy. There are images in the dataset that include photographs of outdoor scenes and leisure photos not at all related to a restaurant. The attribute distribution across images is not uniform, as there are some attributes that occur more frequently than others. Duplicate information can occur as well, as

Table 10.3 Categories: names and frequencies

ID	Label	Relative frequency
0	Good for lunch	0.336
1	Good for dinner	0.497
2	Takes reservations	0.513
3	Outdoor seating	0.502
4	Restaurant is expensive	0.274
5	Has alcohol	0.625
6	Has table service	0.680
7	Ambience is classy	0.286
8	Good for kids	0.619

a consequence of users accidentally uploading the same photo of the same business more than once.

The images are of variable size, ranging from icon-size to 500 × 500, although almost all of them are larger than the required input size of 224 × 224. Figure 10.5 contains examples of restaurant pictures and food items. Approximately 70 % of the pictures of a restaurant are of food items, a good number of these being shots of various items kept on the table. This information can be used for obtaining information on suitability for lunch/dinner, alcohol, or table service. Other categories are more difficult to obtain.

10.6.2 Winners of Round 6 Yelp Challenge

355 Kagglers accepted the challenge of Yelp to predict multiple attribute labels for restaurants based on user-submitted photos. First place was awarded to Dmitrii Tsybulevskii. Thuyen Ngo came in second. We comment on their work below.

First Place, Dmitrii Tsybulevskii

Dmitrii Tsybulevskii took first place in this competition. In order to tackle the multi-label and multi-instance aspects of this problem, he used the *embedded space paradigm* (Sect. 5.3), where each bag is mapped to a single feature vector summarizing its relevant information. To deal with the multi-label component, he used Binary Relevance (BR) and Ensemble of Classifier Chains (ECC) with binary classification methods. His best performing model was the multi-output neural network. This network shares weights for the different label learning tasks and performs better than several BR or ECC neural networks with binary outputs, because it takes into account the multi-label aspect.

Second Place, Thuyen Ngo

Thuyen Ngo ranked in second place in this competition. He used a multilayer perceptron to handle the multiple label and multiple instance aspects at the same time.

For the multi-label part, he used 9 sigmoid units. To address the multi-instance task, he employed a procedure known as the *attention mechanism* in the neural network literature. The idea is to let the network learn by itself how to combine information from many instances. The model is trained using the business-level labels, such that each business represents a training sample. Standard cross entropy is used as the loss function. With limited labeled data, this approach would have badly overfitted the data, since it has more than 2M parameters. To remedy this, Thuyen Ngo used dropout for almost all layers and early stopping to mitigate overfitting.

10.7 Relevant Multi-instance Multi-label Learning Research Directions

As discussed in Sect. 3.5, the inherent features of MIL require a careful study of appropriate distance measures. When MIL is combined with MLL, this topic becomes even more important. Jin et al. [9] proposed an iterative algorithm for MIMLL distance metric learning. Their proposal first estimates the association between instances in a bag and the assigned class labels. Next, it learns a distance metric from the estimated association by means of discriminative analysis. Finally, the learned metric is used to update the association between instances and class labels, which is further used to improve the learning of the metric.

Another relevant area in any learning paradigm is the improvement of the algorithmic efficiency. This task is more pronounced in MIMLL because its hypothesis space expands dramatically, resulting in high complexity and limiting this type of applications. A few studies deal with this problem directly. Huang et al. [8] proposed the MIMLfast approach, which first constructs a low-dimensional subspace shared by all labels and then trains label-specific linear models to optimize the approximated ranking loss via stochastic gradient descent. Ren et al. [20] adapted MIMLfast to perform appropriately in specific classification problems with a small quantity of high-quality data. High-quality data are data where the number of training bags is much less than the number of features.

We also encounter studies that exploit the power of the MIMLL framework by combining it with others. In recent years, many learning methods have been proposed to work with multi-view data by considering the diversity of different views. These views may be obtained from multiple sources or different feature subsets. The learning task can be conducted with abundant information showing a better generalization ability than single-view learning. The combination of multi-view, multi-instance, and multi-label learning has shown a greater flexibility for representing objects. Nguyen et al. [11] proposed a Multimodal Multi-instance Multi-label Latent Dirichlet Allocation (M3LDA), where the model consists of a visual-label part, a textual-label part, and a label topic part that allows to work with discrete views. An extension of this work was carried out by Nguyen et al. [12], presenting the Multi-Instance Multi-Label Mixture (MIMLmix) algorithm, a more efficient model that allows to work with continuous views. Wu et al. [27] modeled the music emotion recognition as a multi-label multi-layer multi-instance multi-view learning problem. Music is formulated as a hierarchical multi-instance structure, where multiple emotion labels correspond to at least one of the instances with multiple views of each layer. To solve this problem, a Hierarchical Music Emotion Recognition model was proposed. Shen et al. [21] combined multi-task multi-label and multi-instance learning and they proposed MTML-MIL, an algorithm based on SVM to leverage both large-scale loosely tagged images and the inter-object correlations for achieving more effective training of a large number of inter-related object classifiers.

Finally, in recent years, we encounter studies that accomplish the specification of novelty detection in the MIMLL setting. Novelty detection is the task of classifying new or unknown data that are not labeled during training and play an important role in machine learning. It is a fundamental requirement of a good classification or identification system, since the test data sometimes contains information about objects that were not known at training time. Contrary to the common assumption in MIMLL that each instance in a bag belongs to one of the known classes, in novelty detection, bags may contain novel-class instances. The goal is to determine, for any given instance in a new bag, whether it belongs to a known class or to a new one. Several works in this line [15, 16, 18] show that novelty detection in the MIMLL setting captures many real-world phenomena and has many potential applications of recognition, such as handwritten digit recognition or letter recognition.

10.8 Summarizing Comments

In solving real-world problems, a good data representation is often more important than having a strong learning algorithm, since a good representation may capture more meaningful information and render the learning task easier to tackle. MIMLL appears as a natural and convenient framework for problems involving complex objects. It provides flexibility in both the input and output space. In this chapter, a description of MIMLL is presented, including a formal definition, applications, and main methods. The recent Yelp dataset challenge is recounted as an illustration of a real-world MIMLL application.

References

1. Abdechiri, M., Faez, K.: Efficacy of utilizing a hybrid algorithmic method in enhancing the functionality of multi-instance multi-label radial basis function neural networks. Appl. Soft Comput. **34**, 788–798 (2015)
2. Briggs, F., Fern, X.Z., Raich, R.: Rank-loss support instance machines for MIML instance annotation. In: Goethals, B. (ed.) Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2012), pp. 534–542. ACM, New York (2012)

3. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc Am. **131**(6), 4640–4650 (2012)
4. Chen, Z., Chi, Z., Fu, H., Feng, D.: Multi-instance multi-label image classification: a neural approach. Neurocomputing **99**, 298–306 (2013)
5. Gibaja, E., Ventura, S.: Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **4**(6), 411–444 (2014)
6. He, J., Gu, H., Wang, Z.: Bayesian multi-instance multi-label learning using Gaussian process prior. Mach. Learn. **88**(1), 273–295 (2012)
7. He, J., Gu, H., Wang, Z.: Multi-instance multi-label learning based on Gaussian process with application to visual mobile robot navigation. Inf. Sci. **190**, 162–177 (2012)
8. Huang, S.J., Zhou, Z.H.: Fast multi-instance multi-label learning. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014), pp. 1868–1874. AAAI Press, Québec (2014)
9. Jin, R., Wang, S., Zhou, Z.H.: Learning a distance metric from multi-instance multi-label data. In: Flynn, P., Mortensen, E. (eds.) Proceedings of 20th International Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 896–902. IEEE, Los Alamitos (2009)
10. Nguyen, N.: A new SVM approach to multi-instance multi-label learning. In: Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) Proceedings of the IEEE International Conference on Data Mining (ICDM 2010), pp. 384–392. Conference Publishing Services, Sydney (2010)
11. Nguyen, C.T., Zhan, D.C., Zhou, Z.H.: Multi-modal image annotation with multi-instance multi-label LDA. In: Rossi, F., Thrun, S. (eds.) Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 1558–1564. AAAI Press, Québec (2013)
12. Nguyen, C.T., Wang, X., Liu, J., Zhou, Z.H.: Labeling complicated objects: multi-view multi-instance multi-label learning. In: Rossi, F., Thrun, S. (eds.) Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 2013–2019. AAAI Press, Québec (2014)
13. Li, Y.X., Ji, S., Kumar, S., Ye, J., Zhou, Z.H.: Drosophila gene expression pattern annotation through multi-instance multi-label learning. IEEE ACM Trans. Comput. Biol. Bioinform. **9**(1), 98–112 (2012)
14. Li, C., Shi, G.: Weights optimization for multi-instance multi-label RBF neural networks using steepest descent method. Neural Comput. Appl. **22**(7), 1563–1569 (2013)
15. Lou, Q., Raich, R., Briggs, F., Fern, X.Z.: Novelty detection under multi-label multi-instance framework. In: Sanei, S., Smaragdis, P., Nandi, A., Ho, A., Larsen, J. (eds.) Proceedings of the International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE, Los Alamitos (2013)
16. Pham, A.T., Raich, R., Fern, X.Z., Arriaga, J.P.: Multi-instance multi-label learning in the presence of novel class instances. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), vol. 3, pp. 2427–2435. Omnipress, Lille Grand Palais (2015)
17. Pham, A.T., Raich, R., Fern, X.Z.: Simultaneous instance annotation and clustering in multi-instance multi-label learning. In: Erdomu, D., Akcakaya, M., Kozat, S., Larsen, J. (eds.) Proceedings of the 25th International Workshop on Machine Learning for Signal Processing (MLSP 2015), pp. 1–6. IEEE, Los Alamitos (2015)
18. Pei, Y., Fern, X.Z.: Constrained instance clustering in multi-instance multi-label learning. Pattern Recogn. Lett. **37**, 107–114 (2014)
19. Peng, L., Xu, X., Wang, G.: An empirical study of automatic image annotation through multi-instance multi-label learning. In: Tan, T., Zhou, M., Wang, Y. (eds.) Proceedings of the IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT 2010), pp. 275–278. Institute of Electrical and Electronics Engineers Inc, Beijing (2010)
20. Ren, D., Ma, L., Zhang, Y., Sunderraman, R., Fox, P.T., Laird, A.R., Turner, J.A., Turner, M.D.: Online biomedical publication classification using multi-instance multi-label algorithms with feature reduction. In: Wang, Y., Lu, J., Howard, N., Hu, X. (eds.) Proceedings of the 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI-CC 2015), pp. 234–241. IEEE, Los Alamitos (2015)

21. Shen, Y., Fan, J.P.: Multi-task multi-label multiple instance learning. J Zhejiang Univ. Sci. C **11**(11), 860–871 (2010)
22. Shen, Y., Peng, J., Feng, X., Fan, J.: Multi-label multi-instance learning with missing object tags. Multimed. Syst. **19**(1), 17–36 (2013)
23. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Tsujii, J., Henderson, J., Pasca, M. (eds.) Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012), pp. 455–465. Association for Computational Linguistics, Stroudsburg (2012)
24. Tong-tong, C., Chan-juan, L., Hai-lin, Z., Shu-sen, Z., Ying, L., Xin-miao, D.: A multi-instance multi-label scene classification method based on multi-kernel fusion. In: Arai, K. (ed.) Proceedings of the Conference on Intelligent Systems (IntelliSys 2015), pp. 782–787. IEEE Service Center, Piscataway (2015)
25. Wu, Q., Ng, M.K., Ye, Y.: Markov-miml: a markov chain-based multi-instance multi-label learning algorithm. Knowl. Inf. Syst. **37**(1), 83–104 (2013)
26. Wu, J.S., Huang, S.J., Zhou, Z.H.: Genome-wide protein function prediction through multi-instance multi-label learning. IEEE ACM Trans. Comput. Biol. Bioinform. **11**(5), 891–902 (2014)
27. Wu, B., Zhong, E., Horner, A., Yang, Q.: Music emotion recognition by multi-label multi-layer multi-instance multi-view learning. In: Cai, Y., Tavanapong, W. (eds.) Proceedings of the 22nd International Conference on Multimedia (MM 2014), pp. 117–126. ACM, New York (2014)
28. Wu, J.S., Hu, H.F., Yan, S.C., Tang, L.H.: Multi-instance multilabel learning with weak-label for predicting protein function in electricigens. Biomed. Res. Int. **2015**, 1–9 (2015)
29. Xu, X.S., Xue, X., Zhou, Z.H.: Ensemble multi-instance multi-label learning approach for video annotation task. In: Sundaram, H., Feng, W.-C., Sebe, N. (eds.) Proceedings of the 19th ACM International Conference on Multimedia (MM 2011), pp. 1153–1156. ACM, New York (2011)
30. Yan, K., Li, Z., Zhang, C.: A New multi-instance multi-label learning approach for image and text classification. Multimed. Tools Appl. **75**(13), 7875–7890 (2015)
31. Yang, S.H., Zha, H., Hu, B.G.: Dirichlet-bernoulli alignment: a generative model for multi-class multi-label multi-instance corpora. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Proceedings of 22nd Conference on Advances in Neural Information Processing Systems (NIPS 2009), pp. 2143–2150. MIT Press, Cambridge (2009)
32. Yang, S.J., Jiang, Y., Zhou, Z.H.: Multi-instance multi-label learning with weak label. In: Rossi, F., Thrun, S. (eds.) Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 1862–1868. AAAI Press, Beijing (2013)
33. Zhang, M.L., Zhou, Z.H.: Multi-label learning by instance differentiation. In: Holte, R.C., Howe, A. (eds.) Proceedings of 22nd Conference on Artificial Intelligence (AAAI 2007), pp. 669–674. AAAI Press, Vancouver (2007)
34. Zhang, M.L., Zhou, Z.H.: M3MIML: a maximum margin method for multi-instance multi-label learning. In: Giannotti, F., Gunopulos, D., Turini, F., Zaniolo, C., Ramakrishnan, N., Wu, X. (eds.) Proceedings of 8th IEEE International Conference on Data Mining (ICDM), pp. 688–697. IEEE, Los Alamitos (2008)
35. Zhang, M.L., Wang, Z.J.: MIMLRBF: RBF neural networks for multi-instance multi-label learning. Neurocomputing **72**(16), 3951–3956 (2009)
36. Zhang, M.L.: A k-nearest neighbor based multi-instance multi-label learning algorithm. In: Gregoire, E. (ed.) Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2010), vol. 2, pp. 207–212. IEEE, Los Alamitos (2010)
37. Zhou, Z.H., Zhang, M.L.: Multi-instance multi-label learning with application to scene classification. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) Proceedings of 19th Conference on Advances in Neural Information Processing Systems (NIPS 2006), pp. 1609–1616. MIT Press, Cambridge (2006)

38. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: MIML: a framework for learning with ambiguous objects. Cornell University Library, pp. 1–57 (2008). arXiv:0808.3231
39. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artif. Intell. **176**(1), 2291–2320 (2012)