多标记学习*

张 敏 灵[1]　　周 志 华[2]

[1] 东南大学计算机科学与工程学院，南京 210096

[2] 南京大学计算机软件新技术国家重点实验室，南京 210093

1. 引言

在机器学习中，传统监督学习（traditional supervised learning）[1]是研究得最多、应用最广泛的一种学习框架。在该框架下，对于真实世界的每一个对象，学习系统在输入空间用一个示例（instance，通常为属性向量）刻画对象的性质，同时在输出空间将示例与反映该对象语义信息的类别标记（label）相关联，这样就得到了一个样本（example）。在拥有了一个较大的样本集合即训练集（training set）之后，学习系统利用某种学习算法学得输入空间（即示例空间）与输出空间（即标记空间）之间的一个映射，基于该映射可以预测未见示例（unseen instance）的类别标记。形式化地说，假设$\mathcal{X}$代表示例空间，$\mathcal{Y}$代表标记空间，则学习系统的任务是从训练集$\{(\boldsymbol{x}_i, y_i) \mid 1 \leq i \leq m\}$中学得函数$f : \mathcal{X} \to \mathcal{Y}$，其中$\boldsymbol{x}_i \in \mathcal{X}$为一个示例而$y_i \in \mathcal{Y}$为示例$\boldsymbol{x}_i$所属的类别标记。在待学习对象具有明确、单一的语义，即对象的类别标记唯一时，上述传统监督学习框架已经取得了巨大的成功。

然而，真实世界的对象往往并不只具有唯一的语义，而是可能具有多义性的。对于图 1(a)所示的一篇关于南非世界杯的新闻报道，既可以认为它属于“体育”这个类别，也可以认为它属于“非洲”这个类别，该报道可能还谈及了本次世界杯对南非在经济层面的影响从而属于“经济”类；再比如，对于图 1(b)所示的图像而言，既可以认为它属于“日落”这个类别，也可以认为它属于“云”、“树木”甚至“乡村”类。这样的例子还有很多，一个基因可能同时具有多种功能如“新陈代谢”、“转录”以及“蛋白质合成”，一首乐曲可能传达了多种信息如“钢琴”、“古典音乐”、“莫扎特”以及“奥地利”，等等。

由上可见，多义性对象由于不再具有唯一的语义，这就使得前述的只考虑明确、单一的语义的传统监督学习框架难以取得好的效果。为了直观地反映多义性对象所具有的多种语义信息，一种很自然的方式就是为该对象显式地赋予一组合适的类别标记，即标记子集。基于上述考虑，作为一种多义性对象学习建模工具，多标记学习（multi-label learning）框架[2][3]由此应运而生。在该框架下，每个对象由一个示例描述，该示例具有多个而不再是唯一的类别标记，学习的目标是将所有合适的类别标记赋予未见示例。

*本文得到国家自然科学基金（60805022）、教育部博士点基金新教师项目（200802941009）以及东南大学引进人才启动基金的资助

(a) 一篇文档　　(b) 一幅图像

图 1 多义性对象的两个例子

早期，多标记学习的研究主要集中于分档分类（text categorization）中遇到的多义性问题[4][5][6][7]。经过近十年来的发展，多标记学习技术已在多媒体内容自动标注[8][9][10]、生物信息学[11][12][13]、Web 挖掘[14][15]、信息检索[16][17][18]、个性化推荐[19][20]等领域得到了广泛应用。据笔者不完全统计，近四年以来（2007 年—2010 年），在与机器学习相关的一流国际会议 ICML、NIPS、ECML/PKDD、KDD、ICDM、IJCAI 以及 AAAI 上，标题部分出现“多标记（multi-label / multilabel）”这一关键词的论文即超过了 30 篇。此外，近两年召开的 ECML/PKDD'09 以及 ICML/COLT'10 均设置了以“Learning from Multi-Label Data”为主题的 Workshop[1]。多标记学习的研究进展也得到了国际机器学习界权威刊物《Machine Learning》的关注，将于近期推出一期以多标记学习为主题的专辑[2]。

总的来说，多标记学习的研究对于多义性对象的学习建模具有十分重要的意义，现已逐渐成为国际机器学习界一个新的研究热点。本章将对多标记学习的研究现状做一个简介，首先给出多标记学习的定义与面临的主要问题，并介绍多标记性能评价指标，然后重点介绍几种具有代表性的多标记学习算法，最后简要讨论多标记学习的拓展研究课题及相关学术资源。

2. 学习框架

2.1 问题定义

假设$\mathcal{X} = \mathbb{R}^d$代表 d 维的示例空间，$\mathcal{Y} = \{y_1, y_2, \cdots, y_q\}$代表包含 q 个类别的标记空间。给定多标记训练集$\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le m\}$，其中$\boldsymbol{x}_i \in \mathcal{X}$为$d$维的属性向量$(x_{i1}, x_{i2}, \cdots, x_{id})^{\mathrm{T}}$，而$Y_i \subset \mathcal{Y}$

[1] MLD'09 - http://lpis.csd.auth.gr/workshops/mld09/ (in conjunction with ECML/PKDD 2009)
MLD'10 - http://cse.seu.edu.cn/conf/MLD10/ (in conjunction with ICML/COLT 2010)
[2] http://mlkd.csd.auth.gr/events/ml2010si.html

为与$\boldsymbol{x}_i$对应的一组类别标记，学习系统的任务是从中学习得到一个多标记分类器$h:\mathcal{X}\to 2^{\mathcal{Y}}$。基于此，对于任一示例$\boldsymbol{x}\in\mathcal{X}$，分类器预测隶属于该示例的类别标记集合为$h(\boldsymbol{x})\subseteq\mathcal{Y}$。

在许多情况下，学习系统的输出往往对应于某个实值函数$f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$，其中$f(\boldsymbol{x},y)$可以看作示例$\boldsymbol{x}$具有类别标记$y$的“置信度（confidence）”。对于给定的示例$\boldsymbol{x}$及其对应的类别标记集合Y，一个成功的学习系统将在隶属于Y的类别标记上输出较大的值，而在不属于Y的类别标记上输出较小的值，即$f(\boldsymbol{x},y')>f(\boldsymbol{x},y'')\ (y'\in Y,y''\notin Y)$成立。此外，实值函数$f(\cdot,\cdot)$还可转化为一个排序函数$rank_f(\cdot,\cdot)$，该排序函数将所有的实值输出$f(\boldsymbol{x},y)\ (y\in\mathcal{Y})$映射到集合$\{1,2,\cdots,q\}$上，使得当$f(\boldsymbol{x},y')>f(\boldsymbol{x},y'')$成立时$rank_f(\boldsymbol{x},y')<rank_f(\boldsymbol{x},y'')$亦成立。

上述的多标记分类器$h(\cdot)$其实可以由实值函数$f(\cdot,\cdot)$转换而来：给定阈值函数$t:\mathcal{X}\to\mathbb{R}$，则$h(\boldsymbol{x})=\{y\mid f(\boldsymbol{x},y)>t(\boldsymbol{x}),\ y\in\mathcal{Y}\}$。换句话说，基于阈值$t(\boldsymbol{x})$学习系统将标记空间二分为“相关（relevant）”标记集合与“无关（irrelevant）”标记集合。阈值函数通常设为“常量函数（constant function）”。值得注意的是，虽然多标记分类器$h(\cdot)$可以通过实值函数$f(\cdot,\cdot)$对标记进行排序（并结合阈值函数）求得，但多标记学习问题和“标记排序（label ranking）”问题[21]在本质上是不同的。

如果限定每个示例只对应一个类别标记，则传统的监督学习框架可以看作多标记学习框架的“特例（degenerated version）”。然而另一方面，多标记学习的“一般性（generality）”使得解决该问题的难度大大增加。总的来看，多标记学习所面临的最大挑战在于其输出空间过大，即输出空间的类别标记集合数将随着标记空间的增大而成指数规模增长。例如，当标记空间具有 20 个类别标记时$(q=20)$，则可能的类别标记集合数将超过一百万（即 2^{20}）。

为了有效应对标记集合空间过大所造成的学习困难，学习系统需要充分利用标记之间的“相关性（correlation）”来辅助学习过程的进行。例如，如果已知一幅图像具有类别标记“狮子”与“草原”，则该图像具有类别标记“非洲”的可能性将会增加；如果已知一篇文档具有类别标记“娱乐”，则该文档同时隶属于类别标记“政治”的可能性将会降低。因此，如何充分利用标记之间的相关性是构造具有强泛化能力多标记学习系统的关键。基于考察标记之间相关性的不同方式，已有的多标记学习问题求解策略大致可以分为以下三类[22]：

a) “一阶（first-order）”策略：该类策略通过逐一考察单个标记而忽略标记之间的相关性，如将多标记学习问题分解为q个独立的二类分类问题，从而构造多标记学习系统。该类方法效率较高且实现简单，但由于其完全忽略标记之间可能存在的相关性，其系统的泛化性能往往较低。

b) “二阶（second-order）”策略：该类策略通过考察两两标记之间的相关性，如相关标记与无关标记之间的排序关系，两两标记之间的交互关系等等，从而构造多标记学习系统。该类方法由于在一定程度上考察了标记之间的相关性，因此其系统泛化性能较优。然而，当真实世界问题中标记之间具有超越二阶的相关性时，该类方法的性能将会受到很大影响。

c) “高阶（high-order）”策略：该类策略通过考察高阶的标记相关性，如处理任一标记对其它所有标记的影响，处理一组随机标记集合的相关性等等，从而构造多标记学习系统。该类方法虽然可以较好地反映真实世界问题的标记相关性，但其模型复杂度往往过高，难以处理大规模学习问题。

由上可见，不同的多标记学习问题求解策略具有各自的优缺点。本章在第 3 节将针对不同的求解策略，介绍几种具有代表性的多标记学习算法。

2.2 评价指标

在多标记学习问题中，由于每个对象可能同时具有多个类别标记，因此传统监督学习中常用的单标记评价指标，如精度（accuracy）、查准率（precision）、查全率（recall）等，无法直接用于多标记学习系统的性能评价。因此，研究者们相继提出了一系列多标记评价指标，总的来看可分为两种类型，即“基于样本”的评价指标（example-based metrics）[4]以及“基于类别”的评价指标（label-based metrics）[23]。

基于样本的多标记评价指标首先衡量分类器在单个测试样本上的分类效果，然后返回其在整个测试集上的“均值（mean value）”作为最终的结果。基于 2.1 节的符号表示，给定多标记分类器$h(\cdot)$以及多标记测试集$\mathcal{S}=\{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq p\}$，其中$Y_i$为隶属于示例$\boldsymbol{x}_i$的相关标记集合。常用的基于样本的多标记评价指标包括：

- *Subset accuracy*:

$$\mathrm{subsetacc}_{\mathcal{S}}(h)=\frac{1}{p}\sum_{i=1}^{p}[\![h(\boldsymbol{x}_i)=Y_i]\!] \tag{1}$$

其中，对于任意的谓词π，当π成立时$[\![\pi]\!]$取值为 1，否则$[\![\pi]\!]$取值为 0。该评价指标用于考察预测的标记集合与真实的标记集合完全吻合的样本占测试集的比例情况。该指标取值越大则系统性能越优，其最优值为$\mathrm{subsetacc}_{\mathcal{S}}(h)=1$。值得注意的是，当标记空间中包含大量类别标记（$q$ 很大）时，学习系统往往难以给出与真实的标记集合完全吻合的预测，此时该评价指标的取值将会很低。

- *Hamming loss*:

$$\mathrm{hloss}_{\mathcal{S}}(h)=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}\left|h(\boldsymbol{x}_i)\Delta Y_i\right| \tag{2}$$

其中，算子Δ用于度量两个集合之间的“对称差（symmetric difference）”，算子$|\cdot|$用于返回集合的“势（cardinality）”。该评价指标用于考察样本在单个标记上的误分类情况，即相关标记未出现在预测的标记集合中或无关标记出现在预测的标记集合中。该指标取值越小则系统性能越优，其最优值为$\mathrm{hloss}_{\mathcal{S}}(h)=0$。值得注意的是，当$\mathcal{S}$中的每个样本仅含有一个类别标记时，hamming loss 的取

值即为传统分类误差的$2/q$倍。

- *One-error*:

$$\text{one-error}_{\mathcal{S}}(h) = \frac{1}{p}\sum_{i=1}^{p}[\![\,[\arg\max_{y\in\mathcal{Y}} f(\boldsymbol{x}_i, y)] \notin Y_i\,]\!] \tag{3}$$

其中，$f(\cdot,\cdot)$为与多标记分类器$h(\cdot)$对应的实值函数。该评价指标用于考察在样本的类别标记排序序列中，序列最前端的标记不属于相关标记集合的情况。该指标取值越小则系统性能越优，其最优值为$\text{one-error}_{\mathcal{S}}(h) = 0$。值得注意的是，当$\mathcal{S}$中的每个样本仅含有一个类别标记时，one-error 即为传统的分类误差。

- *Coverage*:

$$\text{coverage}_{\mathcal{S}}(h) = \frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i} rank_f(\boldsymbol{x}_i, y) - 1 \tag{4}$$

其中，$rank_f(\cdot,\cdot)$为与实值函数$f(\cdot,\cdot)$对应的排序函数。该评价指标用于考察在样本的类别标记排序序列中，覆盖所有相关标记所需的搜索深度情况。该指标取值越小则系统性能越优，其最优值为$\text{coverage}_{\mathcal{S}}(h) = \frac{1}{p}\sum_{i=1}^{p}|Y_i| - 1$。

- *Ranking loss*:

$$\text{rloss}_{\mathcal{S}}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}|\{(y', y'') \mid f(\boldsymbol{x}_i, y') \le f(\boldsymbol{x}_i, y''),\ (y', y'') \in Y_i \times \bar{Y}_i\}| \tag{5}$$

其中，$\bar{Y}_i$为集合Y_i在标记空间$\mathcal{Y}$中的“补集（complementary set）”。该评价指标用于考察在样本的类别标记排序序列中出现排序错误的情况，即无关标记在排序序列中位于相关标记之前。该指标取值越小则系统性能越优，其最优值为$\text{rloss}_{\mathcal{S}}(h) = 0$。

- *Average precision*:

$$\text{avgprec}_{\mathcal{S}}(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|\{y' \mid rank_f(\boldsymbol{x}, y') \le rank_f(\boldsymbol{x}_i, y),\ y' \in Y_i\}|}{rank_f(\boldsymbol{x}_i, y)} \tag{6}$$

该评价指标用于考察在样本的类别标记排序序列中，排在相关标记之前的标记仍为相关标记的情况。该指标取值越大则系统性能越优，其最优值为$\text{avgprec}_{\mathcal{S}}(h) = 1$。值得注意的是，该指标最先出现于信息检索领域，用于度量给定查询下检索系统返回文档的排序性能[24]。

与基于样本的多标记评价指标不同，基于类别的多标记评价指标首先衡量分类器在单个类别上对应的“二类分类（binary classification）”效果，然后返回其在所有类别上的“均值（macro-/micro-averaged value）”作为最终的结果。给定多标记测试集$\mathcal{S} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le p\}$，对于第$j$个类别

y_j $(1 \leq j \leq q)$而言，分类器$h(\cdot)$在该类别上的二类分类性能可由如下四个统计量进行刻画：

- TP_j（“真”正例的个数，#true positive instances）：

$$TP_j = |\{\boldsymbol{x}_i \mid y_j \in Y_i \bigwedge y_j \in h(\boldsymbol{x}_i),\ (\boldsymbol{x}_i, Y_i) \in \mathcal{S}\}| \tag{7}$$

- FP_j（“伪”正例的个数，#false positive instances）：

$$FP_j = |\{\boldsymbol{x}_i \mid y_j \notin Y_i \bigwedge y_j \in h(\boldsymbol{x}_i),\ (\boldsymbol{x}_i, Y_i) \in \mathcal{S}\}| \tag{8}$$

- TN_j（“真”负例的个数，#true negative instances）：

$$TN_j = |\{\boldsymbol{x}_i \mid y_j \notin Y_i \bigwedge y_j \notin h(\boldsymbol{x}_i),\ (\boldsymbol{x}_i, Y_i) \in \mathcal{S}\}| \tag{9}$$

- FN_j（“伪”负例的个数，#false negative instances）：

$$FN_j = |\{\boldsymbol{x}_i \mid y_j \in Y_i \bigwedge y_j \notin h(\boldsymbol{x}_i),\ (\boldsymbol{x}_i, Y_i) \in \mathcal{S}\}| \tag{10}$$

据式(7)-式(10)可知，$TP_j + FP_j + TN_j + FN_j = p$成立。值得注意的是，绝大部分二类分类性能指标均可由以上四个统计量导出，例如：

$$\text{Accuracy} = B(TP_j, FP_j, TN_j, FN_j) = \frac{TP_j + TN_j}{TP_j + FP_j + TN_j + FN_j} \tag{11}$$

$$\text{Precision} = B(TP_j, FP_j, TN_j, FN_j) = \frac{TP_j}{TP_j + FP_j} \tag{12}$$

$$\text{Recall} = B(TP_j, FP_j, TN_j, FN_j) = \frac{TP_j}{TP_j + FN_j} \tag{13}$$

基于此，令$B(TP_j, FP_j, TN_j, FN_j)$代表由所定义的统计量求得的某种二类分类性能指标，则基于类别的多标记评价指标可采用如下两种方式获得：

- *Macro-averaging*:

$$B_{\text{macro}} = \frac{1}{q} \sum_{j=1}^{q} B(TP_j, FP_j, TN_j, FN_j) \tag{14}$$

- *Micro-averaging*:

$$B_{\text{micro}} = B\left(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\right) \tag{15}$$

其中，macro-averaging 首先基于统计量求得在各个类上的分类性能，然后再将所有类上的均值作为最终结果，其基本思想是为各个类赋予相同的权重。相应地，micro-averaging 首先将各个类上的统计量相加，然后再将求得的分类性能作为最终结果，其基本思想是为各个样本赋予相同的权重。

总的来看，已有的各种基于样本的或者基于类别的多标记评价指标是从不同的侧面来衡量学习系统的泛化性能。目前并不存在适用于所有问题的“通用的（general-purpose）”多标记评价指标，其选择依赖于具体的学习任务。例如，对于“分类（classification）”任务而言，采用基于样本的评价指标如 hamming loss 可能比较合适；而对于“检索（retrieval）”任务而言，采用基于类别的评价指标如 micro-averaged precision 可能比较合适。除此之外，各评价指标之间的关系尚不明确，如优化其中一些指标是否意味着同时优化其他一些指标等，关于这方面的研究工作目前还比较少[25]。

3. 学习算法

3.1 算法分类

一般而言，算法研究是机器学习研究的核心课题，这一点对于多标记学习而言也不例外。目前已经涌现出了大量的多标记学习算法，总的来看大致可以分为两类：

a) “问题转换（problem transformation）”方法：该类方法的基本思想是通过对多标记训练样本进行处理，将多标记学习问题转换为其它已知的学习问题进行求解。代表性学习算法有一阶方法 Binary Relevance[8]，该方法将多标记学习问题转化为“二类分类（binary classification）”问题求解；二阶方法 Calibrated Label Ranking[26]，该方法将多标记学习问题转化为“标记排序（label ranking）”问题求解；高阶方法 Random *k*-labelsets[23]，该方法将多标记学习问题转化为“多类分类（multi-class classification）”问题求解。

b) “算法适应（algorithm adaptation）”方法：该类方法的基本思想是通过对常用监督学习算法进行改进，将其直接用于多标记数据的学习。代表性学习算法有一阶方法 ML-*k*NN[27]，该方法将“惰性学习（lazy learning）”算法 *k* 近邻进行改造以适应多标记数据；二阶方法 Rank-SVM[11]，该方法将“核学习（kernel learning）”算法 SVM 进行改造以适应多标记数据；高阶方法 LEAD[22]，该方法将“贝叶斯学习（Bayes learning）算法”Bayes 网络进行改造以适应多标记数据。

换句话说，问题转换方法的核心是“改造数据适应算法（fit data to algorithm）”，本章 3.1 小节将介绍前述的三种基于该方法的代表性算法；算法适应方法的核心是“改造算法适应数据（fit algorithm to data）”，本章 3.2 小节将介绍前述的三种基于该方法的代表性算法。

3.2 “问题转换”算法

3.2.1 Binary Relevance

该算法的基本思想是将多标记学习问题转化为 q 个独立的二类分类问题，其中每个二类分类问题对应于标记空间$\mathcal{Y}$中的一个类别标记[8]。

基于 2.1 节的符号表示，给定多标记训练集$\mathcal{D}=\{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$，其中$Y_i$为隶属于示例$\boldsymbol{x}_i$

的相关标记集合。具体来说，对于第j个类别y_j $(1 \le j \le q)$而言，Binary Relevance 算法首先构造与该类别对应的二类训练集：

$$\mathcal{D}_j = \{(\boldsymbol{x}_i, \phi(Y_i, y_j)) \mid 1 \le i \le m\} \tag{16}$$

$$\text{where} \quad \phi(Y_i, y_j) = \begin{cases} +1, & \text{if } y_j \in Y_i \\ -1, & \text{otherwise} \end{cases}$$

基于此，Binary Relevance 算法采用某种二类学习算法$\mathcal{B}$训练二类分类器$g_j : \mathcal{X} \to \mathbb{R}$，即$g_j \leftarrow \mathcal{B}(\mathcal{D}_j)$。由此可见，对于任一多标记样本$(\boldsymbol{x}_i, Y_i)$，示例$\boldsymbol{x}_i$将参与 q 个二类分类器的学习。其中，对于相关标记$y_j \in Y_i$而言，$\boldsymbol{x}_i$在构造二类分类器$g_j(\cdot)$时对应于正例；对于无关标记$y_j \in \bar{Y}_i$而言，$\boldsymbol{x}_i$在构造二类分类器$g_j(\cdot)$时对应于反例。该训练策略亦称为“交叉训练（cross-training）”法[8]。

在测试阶段，对于未见示例$\boldsymbol{x}$，Binary Relevance 算法通常采用如下方式预测其类别标记集合Y：

$$Y = \{y_j \mid g_j(\boldsymbol{x}) > 0, 1 \le j \le q\} \tag{17}$$

值得注意的是，当所有二类分类器的输出均为负值时，将会导致算法预测的标记集合Y为空。为了避免这种情况的发生，可以采用如下的 T-Criterion 准则[8]来进行预测：

$$Y = \{y_j \mid g_j(\boldsymbol{x}) > 0, 1 \le j \le q\} \bigcup \{y_{j^*} \mid j^* = \arg\max_{1<j<q} g_j(\boldsymbol{x})\} \tag{18}$$

此时，当所有二类分类器输出为负时，预测的标记集合Y中将含有输出值“最大（least negative）”的类别标记。除了上述的 T-Criterion 准则之外，Boutell 等人[8]还给出了其它一些基于各二类分类器输出确定测试样本标记集合的准则，具体细节可参见相应文献。

3.2.2 Calibrated Label Ranking

该算法的基本思想是将多标记学习问题转化为标记排序问题，其中标记排序采用“成对比较（pairwise comparison）”的方式实现[26]。

对于具有 q 个类别的标记空间而言，针对每一个可能的标记配对(y_j, y_k) $(1 \le j < k \le q)$，采用成对比较的方式将产生共计$q(q-1)/2$个二类分类器。具体来说，对于标记配对(y_j, y_k)而言，成对比较法首先构造与该配对对应的二类训练集：

$$\mathcal{D}_{jk} = \{(\boldsymbol{x}_i, \psi(Y_i, y_j, y_k)) \mid \phi(Y_i, y_j) \ne \phi(Y_i, y_k),\ 1 \le i \le m\} \tag{19}$$

$$\text{where} \quad \psi(Y_i, y_j, y_k) = \begin{cases} +1, & \text{if } \phi(Y_i, y_j) = +1 \text{ and } \phi(Y_i, y_k) = -1 \\ -1, & \text{if } \phi(Y_i, y_j) = -1 \text{ and } \phi(Y_i, y_k) = +1 \end{cases}$$

其中，算子$\phi(\cdot,\cdot)$的定义如式(16)所示。基于此，成对比较法采用某种二类学习算法$\mathcal{B}$训练二类分类

器$g_{jk}:\mathcal{X}\to\mathbb{R}$，即$g_{jk}\leftarrow\mathcal{B}(\mathcal{D}_{jk})$。由此可见，对于任一多标记样本$(\boldsymbol{x}_i,Y_i)$，示例$\boldsymbol{x}_i$将参与$|Y_i||\bar{Y}_i|$个二类分类器的学习。其中，对于$(y_j,y_k)\in Y_i\times\bar{Y}_i\ (j<k)$的情况而言，$\boldsymbol{x}_i$在构造二类分类器$g_{jk}(\cdot)$时对应于正例；对于$(y_j,y_k)\in\bar{Y}_i\times Y_i\ (j<k)$的情况而言，$\boldsymbol{x}_i$在构造二类分类器$g_{jk}(\cdot)$时对应于反例。

在测试阶段，对于未见示例$\boldsymbol{x}$，Calibrated Label Ranking 算法首先将其提交给已训练的$q(q-1)/2$个二类分类器，得到该示例在各个类别标记上的“投票（votes）”：

$$\zeta(\boldsymbol{x},y_j)=\sum_{k=1}^{j-1}[\![g_{kj}(\boldsymbol{x})\le 0]\!]+\sum_{k=j+1}^{q}[\![g_{jk}(\boldsymbol{x})>0]\!]\qquad(1\le j\le q)\tag{20}$$

据式(20)可知，$\sum_{j=1}^{q}\zeta(\boldsymbol{x},y_j)=q(q-1)/2$成立。基于 2.1 节的符号表示，令$f(\boldsymbol{x},y_j)=\zeta(\boldsymbol{x},y_j)$，则可根据相应的排序函数$rank_f(\cdot,\cdot)$对标记空间$\mathcal{Y}$中的所有类别标记实现排序。当$f(\boldsymbol{x},y')=f(\boldsymbol{x},y'')$时，标记$y'$与$y''$的相对排序位置随机确定。

值得注意的是，利用成对比较法虽然可以得到函数$f(\cdot,\cdot)$，实现对所有标记的排序。但如 2.1 节所示，为了得到最终的多标记分类器$h(\cdot)$，仍需确定相应的阈值函数$t(\cdot)$，从而将标记的排序序列“二分（bipartition）”为相关标记集合与无关标记集合。为了在成对比较的框架下实现该目标，Calibrated Label Ranking 算法为每个多标记样本$(\boldsymbol{x}_i,Y_i)$加入一个“虚拟标记（virtual label）”y_{V}，该虚拟标记的作用是在$\boldsymbol{x}_i$的相关标记集合与无关标记集合之间加入一个“人工分割点（artificial splitting point）”。换句话说，在标记排序序列中，虚拟标记应位于所有相关标记之后，并位于所有无关标记之前。

此时，针对每一个新的标记配对$(y_j,y_{\mathrm{V}})\ (1\le j\le q)$，Calibrated Label Ranking 算法将在原有的$q(q-1)/2$个二类分类器基础上，额外训练q个二类分类器。具体来说，对于标记配对(y_j,y_{V})而言，首先构造与该配对对应的二类训练集：

$$\mathcal{D}_{j\mathrm{V}}=\{(\boldsymbol{x}_i,\varphi(Y_i,y_j,y_{\mathrm{V}}))\mid 1\le i\le m\}\tag{21}$$

$$\text{where}\quad\varphi(Y_i,y_j,y_{\mathrm{V}})=\begin{cases}+1, & \text{if } y_j\in Y_i\\ -1, & \text{otherwise}\end{cases}$$

基于此，Calibrated Label Ranking 算法采用二类学习算法$\mathcal{B}$训练与虚拟标记对应的二类分类器$g_{j\mathrm{V}}:\mathcal{X}\to\mathbb{R}$，即$g_{j\mathrm{V}}\leftarrow\mathcal{B}(\mathcal{D}_{j\mathrm{V}})$。基于新求得的二类分类器$g_{j\mathrm{V}}(\cdot)\ (1\le j\le q)$，可在式(20)的基础上更新未见示例$\boldsymbol{x}$在各个类别标记上的投票：

$$\zeta^*(\boldsymbol{x},y_j)=\zeta(\boldsymbol{x},y_j)+[\![g_{j\mathrm{V}}(\boldsymbol{x})>0]\!]\qquad(1\le j\le q)\tag{22}$$

此外，进一步计算未见示例$\boldsymbol{x}$在虚拟标记上的投票：

$$\zeta^*(\boldsymbol{x},y_{\mathrm{V}})=\sum_{j=1}^{q}[\![g_{j\mathrm{V}}(\boldsymbol{x})\le 0]\!]\tag{23}$$

基于 2.1 节的符号表示，令$f(\boldsymbol{x},y_j)=\zeta^*(\boldsymbol{x},y_j)$且$t(\boldsymbol{x})=\zeta^*(\boldsymbol{x},y_{\mathrm{V}})$，可得到所需的多标记分类器：

$$h(\boldsymbol{x}) = \{y_j \mid \zeta^*(\boldsymbol{x}, y_j) > \zeta^*(\boldsymbol{x}, y_{\mathrm{V}}),\ 1 \le j \le q\} \tag{24}$$

值得注意的是，对照式(21)与式(16)的定义，训练集$\mathcal{D}_{j\mathrm{V}}$即为 Binary Relevance 算法所使用的训练集$\mathcal{D}_j$。因此，Calibrated Label Ranking 算法可以看作是在常规标记配对求得的$q(q-1)/2$个二类分类器基础上，进一步引入 Binary Relevance 算法求得的 q 个二类分类器，以辅助学习任务的完成[26]。

3.2.3 Random *k*-Labelsets

该算法的基本思想是将多标记学习问题转化为多类分类问题的“集成（ensemble）”，集成中的每一个基分类器对应于标记空间$\mathcal{Y}$的一个随机子集，并采用“Label Powerset”的方式进行构造[23]。

简单地说，Label Powerset（简记为 LP）是一种直观地将多标记学习问题转化为多类分类问题的方法。对于包含 q 个类别的标记空间$\mathcal{Y}$而言，给定多标记训练集$\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le m\}$，我们可以将训练集中出现的每一种标记组合看作一个“新类（new class）”。不失一般性，设$\sigma_{\mathcal{Y}} : 2^{\mathcal{Y}} \to \mathbb{N}$为标记空间$\mathcal{Y}$的“幂空间”至自然数空间的“单射函数（injective function）”，而$\sigma_{\mathcal{Y}}^{-1}$为与$\sigma_{\mathcal{Y}}$对应的“逆函数（inverse function）”。首先，LP 方法将原始的多标记训练集$\mathcal{D}$转化为如下的多类（单标记）训练集：

$$\mathcal{D}_{\mathcal{Y}}^{\dagger} = \{(\boldsymbol{x}_i, \sigma_{\mathcal{Y}}(Y_i)) \mid 1 \le i \le m\} \tag{25}$$

其中，数据集$\mathcal{D}_{\mathcal{Y}}^{\dagger}$中含有新类：

$$\Lambda(\mathcal{D}_{\mathcal{Y}}^{\dagger}) = \{\sigma_{\mathcal{Y}}(Y_i) \mid 1 \le i \le m\} \tag{26}$$

显然，$\left|\Lambda(\mathcal{D}_{\mathcal{Y}}^{\dagger})\right| \le \min(m, 2^{|\mathcal{Y}|})$成立。基于此，LP 方法采用某种多类学习算法$\mathcal{M}$训练多类分类器$g_{\mathcal{Y}}^{\dagger} : \mathcal{X} \to \Lambda(\mathcal{D}_{\mathcal{Y}}^{\dagger})$，即$g_{\mathcal{Y}}^{\dagger} \leftarrow \mathcal{M}(\mathcal{D}_{\mathcal{Y}}^{\dagger})$。由此可见，对于任一多标记样本$(\boldsymbol{x}_i, Y_i)$，示例$\boldsymbol{x}_i$的标记集合首先被映射为一个新类$\sigma_{\mathcal{Y}}(Y_i)$，然后参与多类分类器的学习。

在测试阶段，对于未见示例$\boldsymbol{x}$，LP 方法采用如下方式预测其类别标记集合Y：

$$Y = \sigma_{\mathcal{Y}}^{-1}\left(g_{\mathcal{Y}}^{\dagger}(\boldsymbol{x})\right) \tag{27}$$

值得注意的是，虽然上述 LP 方法可以将多标记学习问题转化为多类分类问题进行求解，但是该方法存在两个主要缺陷。首先，由式(26)及式(27)可知，LP 仅能预测在训练集中出现过的类别标记集合（即$\{Y_i \mid 1 \le i \le m\}$），对于其真实标记集合在训练集中未出现的测试示例无法正确预测；其次，当标记空间$\mathcal{Y}$较大时，往往会导致新类集合$\Lambda(\mathcal{D}_{\mathcal{Y}}^{\dagger})$过大，从而导致部分新类在$\mathcal{D}_{\mathcal{Y}}^{\dagger}$中的训练样本不足且多类分类器的训练复杂度过高。

为了充分发挥 LP 方法简单直观的优势并同时克服其存在的缺陷，Tsoumakas 与 Vlahavas 提出了 Random *k*-Labelsets 算法，结合“集成学习（ensemble learning）”技术与 LP 方法求解多标记学习问

题。其算法核心是每次仅针对一个随机“*k*-标记集（*k*-labelsets）”调用 LP 方法，并将多次调用所得的多类分类器进行集成以得到最终的输出。其中，“*k*-标记集”是标记空间$\mathcal{Y}$的一个子集，包含 *k* 个类别标记。

设$\mathcal{Y}^k$为标记空间$\mathcal{Y}$的所有“*k*-标记集”所构成的集合，其中$|\mathcal{Y}^k| = \binom{q}{k}$。不失一般性，记$\mathcal{Y}^k$中的第 *l* 个“*k*-标记集”为$\mathcal{Y}^k(l)$，则$\mathcal{Y}^k(l) \subseteq \mathcal{Y}$，$|\mathcal{Y}^k(l)| = k$，$1 \le l \le \binom{q}{k}$。采用与式(25)相同的符号表示，对于$\mathcal{Y}^k(l)$而言，在调用 LP 方法时首先构造与之对应的训练集：

$$\mathcal{D}^{\dagger}_{\mathcal{Y}^k(l)} = \left\{ \left(\boldsymbol{x}_i \,,\, \sigma_{\mathcal{Y}^k(l)}(Y_i \cap \mathcal{Y}^k(l)) \right) \;\middle|\; 1 \le i \le m \right\} \tag{28}$$

其中，数据集$\mathcal{D}^{\dagger}_{\mathcal{Y}^k(l)}$中含有新类：

$$\Lambda(\mathcal{D}^{\dagger}_{\mathcal{Y}^k(l)}) = \{ \sigma_{\mathcal{Y}^k(l)}(Y_i \cap \mathcal{Y}^k(l)) \mid 1 \le i \le m \} \tag{29}$$

基于此，采用某种多类学习算法$\mathcal{M}$训练多类分类器$g^{\dagger}_{\mathcal{Y}^k(l)} : \mathcal{X} \to \Lambda(\mathcal{D}^{\dagger}_{\mathcal{Y}^k(l)})$，即$g^{\dagger}_{\mathcal{Y}^k(l)} \leftarrow \mathcal{M}(\mathcal{D}^{\dagger}_{\mathcal{Y}^k(l)})$。

此外，假设 Random *k*-Labelsets 算法所考察的集成大小为 *n*，即针对 *n* 个随机“*k*-标记集”$\{\mathcal{Y}^k(l_1), \cdots, \mathcal{Y}^k(l_n)\}$分别调用 LP 方法，得到相应的多类分类器$\{g^{\dagger}_{\mathcal{Y}^k(l_1)}, \cdots, g^{\dagger}_{\mathcal{Y}^k(l_n)}\}$。基于此，对于未见示例$\boldsymbol{x}$，针对各个类别标记计算如下统计量：

$$\tau(\boldsymbol{x}, y_j) = \sum\nolimits_{r=1}^{n} [\![y_j \in \mathcal{Y}^k(l_r)]\!] \qquad (1 \le j \le q) \tag{30}$$

$$\mu(\boldsymbol{x}, y_j) = \sum\nolimits_{r=1}^{n} \left[\!\!\left[y_j \in \sigma^{-1}_{\mathcal{Y}^k(l_r)} \left(g^{\dagger}_{\mathcal{Y}^k(l_r)}(\boldsymbol{x}) \right) \right]\!\!\right] \qquad (1 \le j \le q) \tag{31}$$

其中，$\tau(\boldsymbol{x}, y_j)$用于统计基于集成在类别标记$y_j$上的最大投票数，而$\mu(\boldsymbol{x}, y_j)$用于统计基于集成在类别标记$y_j$上的实际投票数。基于 2.1 节的符号表示，令$f(\boldsymbol{x}, y_j) = \mu(\boldsymbol{x}, y_j)/\tau(\boldsymbol{x}, y_j)$且$t(\boldsymbol{x}) = 0.5$，则可得到所需的多标记分类器：

$$h(\boldsymbol{x}) = \{ y_j \mid \mu(\boldsymbol{x}, y_j)/\tau(\boldsymbol{x}, y_j) > 0.5,\ 1 \le j \le q \} \tag{32}$$

换句话说，当集成在y_j上的实际投票数超过最大投票数的半数时，该标记即被认为是未见示例的相关标记。一般而言，对于由 *n* 个“*k*-标记集”生成的集成而言，每个类别标记所能得到的最大投票数的平均值为nk/q。Random *k*-Labelsets 算法推荐的默认设置为$k = 3, n = 2q$，此时各类别标记最大投票数的平均值为 6。

本小节分别介绍了三种具有代表性的“问题转换”类型的多标记学习算法，即一阶方法 Binary Relevance[8]、二阶方法 Calibrated Label Ranking[26]以及高阶方法 Random *k*-Labelsets[23]。除此之外，目前还存在其他一些基于“问题转换”的一阶[4][28]、二阶[4][29]以及高阶[30][31][32][33][34]多标记学习算法。限于篇幅，这里不再做一一介绍。

3.3 “算法适应”算法

3.3.1 ML-*k*NN

该算法的基本思想是采用“*k* 近邻（*k*-nearest neighbors）”分类准则，统计近邻样本的类别标记信息，通过“最大化后验概率（maximum a posteriori，简记为 MAP）”的方式推理未见示例的标记集合[27]。

给定多标记训练集$\mathcal{D}=\{(\boldsymbol{x}_i,Y_i)\mid 1\le i\le m\}$以及未见示例$\boldsymbol{x}$，假设$\mathcal{N}(\boldsymbol{x})$代表$\boldsymbol{x}$在训练集中的 *k* 个近邻样本构成的集合。对于第j个类别y_j $(1\le j\le q)$而言，ML-*k*NN 算法将计算如下的统计量：

$$C_j=\sum\nolimits_{(\boldsymbol{x}^*,Y^*)\in\mathcal{N}(\boldsymbol{x})}[\![y_j\in Y^*]\!] \tag{33}$$

由上可知，C_j统计了$\mathcal{N}(\boldsymbol{x})$中将$y_j$作为其相关标记的样本个数。

进一步地，设H_j代表$\boldsymbol{x}$具有类别标记y_j这一事件，$\mathbb{P}(H_j\mid C_j)$代表当$\mathcal{N}(\boldsymbol{x})$中有$C_j$个样本具有类别标记$y_j$时，$H_j$成立的后验概率。相应的，$\mathbb{P}(\neg H_j\mid C_j)$代表当$\mathcal{N}(\boldsymbol{x})$中有$C_j$个样本具有类别标记$y_j$时，$H_j$不成立的后验概率。基于 2.1 节的符号表示，令$f(\boldsymbol{x},y_j)=\mathbb{P}(H_j\mid C_j)/\mathbb{P}(\neg H_j\mid C_j)$且$t(\boldsymbol{x})=0.5$，可得到所需的多标记分类器：

$$h(\boldsymbol{x})=\{y_j\mid \mathbb{P}(H_j\mid C_j)/\mathbb{P}(\neg H_j\mid C_j)>0.5,\ 1\le j\le q\} \tag{34}$$

换句话说，当后验概率$\mathbb{P}(H_j\mid C_j)$大于后验概率$\mathbb{P}(\neg H_j\mid C_j)$时，即将标记$y_j$赋予示例$\boldsymbol{x}$。基于贝叶斯定理，函数$f(\boldsymbol{x},y_j)$可重写为：

$$f(\boldsymbol{x},y_j)\ =\ \frac{\mathbb{P}(H_j\mid C_j)}{\mathbb{P}(\neg H_j\mid C_j)}\ =\ \frac{\mathbb{P}(H_j)\cdot\mathbb{P}(C_j\mid H_j)}{\mathbb{P}(\neg H_j)\cdot\mathbb{P}(C_j\mid \neg H_j)} \tag{35}$$

其中，$\mathbb{P}(H_j)$与$\mathbb{P}(\neg H_j)$分别代表事件H_j成立与不成立的先验概率，$\mathbb{P}(C_j\mid H_j)$与$\mathbb{P}(C_j\mid \neg H_j)$分别代表事件$H_j$成立与不成立时，$\mathcal{N}(\boldsymbol{x})$中有$C_j$个样本具有类别标记$y_j$的条件概率。

值得注意的是，上述先验概率以及条件概率可基于训练集通过“频率计数（frequency counting）”的方式进行估计。具体来说，先验概率可以通过如下方式估计而得：

$$\mathbb{P}(H_j)=\frac{s+\sum_{i=1}^{m}[\![y_j\in Y_i]\!]}{s\times 2+m};\quad \mathbb{P}(\neg H_j)=1-\mathbb{P}(H_j)\qquad (1\le j\le q) \tag{36}$$

其中，“平滑（smoothing）”参数 *s* 用以控制“均匀分布（uniform prior）”在概率估计时的权重，通常 *s* 设置为 1 对应于 Laplace 平滑。

与先验概率的估计不同，条件概率的估计过程要相对复杂一些。对于第j个类别y_j $(1\le j\le q)$而言，ML-*k*NN 算法首先确定两个数组κ_j以及$\tilde{\kappa}_j$，其中每个数组各含有如下 *k*+1 个元素：

$$\kappa_j[r] = \sum_{i=1}^{m} [\![y_j \in Y_i]\!] \cdot [\![\delta_j(\boldsymbol{x}_i) = r]\!] \qquad (0 \le r \le k) \tag{37}$$

$$\tilde{\kappa}_j[r] = \sum_{i=1}^{m} [\![y_j \notin Y_i]\!] \cdot [\![\delta_j(\boldsymbol{x}_i) = r]\!] \qquad (0 \le r \le k) \tag{38}$$

$$\text{where} \quad \delta_j(\boldsymbol{x}_i) = \sum\nolimits_{(\boldsymbol{x}^*, Y^*) \in \mathcal{N}(\boldsymbol{x}_i)} [\![y_j \in Y^*]\!] \tag{39}$$

其中，与式(33)类似，式(39)中定义的$\delta_j(\boldsymbol{x}_i)$统计了第 *i* 个训练样本的 *k* 近邻中，将y_j作为其相关标记的近邻个数。相应地，$\kappa_j[r]$统计了具有标记y_j且其 *k* 近邻中恰好有 *r* 个近邻具有标记y_j的训练样本个数，$\tilde{\kappa}_j[r]$统计了不具有标记y_j且其 *k* 近邻中恰好有 *r* 个近邻具有标记y_j的训练样本个数。基于此，条件概率可以通过如下方式估计而得：

$$\mathbb{P}(C_j \mid H_j) = \frac{s + \kappa_j[C_j]}{s \times (k+1) + \sum_{r=0}^{k} \kappa_j[r]} \qquad (1 \le j \le q,\ 0 \le C_j \le k) \tag{40}$$

$$\mathbb{P}(C_j \mid \neg H_j) = \frac{s + \tilde{\kappa}_j[C_j]}{s \times (k+1) + \sum_{r=0}^{k} \tilde{\kappa}_j[r]} \qquad (1 \le j \le q,\ 0 \le C_j \le k) \tag{41}$$

此时，将所得先验概率(式(36))以及条件概率(式(40)与式(41))代入式(35)，即可基于式(34)得到所需的多标记分类器。

值得一提的是，虽然 ML-*k*NN 算法采用了一阶策略来求解多标记学习问题，即在模型构建过程中忽略标记之间的相互影响。然而，基于该算法的基本思想，可以方便地将其扩展至高阶策略予以实现。例如，在确定事件H_j是否成立时，可以根据$C_1, C_2, \cdots, C_q$中蕴含的信息来进行 MAP 推理（而非仅仅考察C_j的取值），即$f(\boldsymbol{x}, y_j) = \mathbb{P}(H_j \mid C_1, C_2, \cdots, C_q) / \mathbb{P}(\neg H_j \mid C_1, C_2, \cdots, C_q)$。近期，德国学者 E. Hüllermeier 教授指导学生 W. Cheng 沿上述思路专文对 ML-*k*NN 算法进行改进，该论文获 2009 年欧洲机器学习会议最佳学生论文奖并被推荐到权威期刊《Machine Learning》发表[35]。

3.3.2 Rank-SVM

该算法的基本思想是采用“最大化间隔（maximum margin）”策略，定义一组线性分类器以最小化式(5)所示的 *ranking loss* 评价指标，并通过引入“核技巧（kernel trick）”处理非线性分类问题[11]。

设学习系统由 *q* 个线性分类器$\mathbf{W} = \{(\boldsymbol{w}_j, b_j) \mid 1 \le j \le q\}$组成，其中$w_j \in \mathbb{R}^d$为与第 *j* 类对应的“权值向量（weight vector）”，而$b_j \in \mathbb{R}$为与第 *j* 类对应的“偏置（bias）”。基于 2.1 节的符号表示，令$f(\boldsymbol{x}, y_j) = \langle \boldsymbol{w}_j, \boldsymbol{x} \rangle + b_j$，算子$\langle \cdot, \cdot \rangle$返回向量内积。给定多标记训练集$\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le m\}$，Rank-SVM 算法首先按如下方式定义学习系统在样本$(\boldsymbol{x}_i, Y_i)$上的分类间隔：

$$\min_{(y_j, y_k) \in Y_i \times \bar{Y}_i} \frac{\langle \boldsymbol{w}_j - \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + b_j - b_k}{\| \boldsymbol{w}_j - \boldsymbol{w}_k \|} \tag{42}$$

其中，对于相关标记y_j以及无关标记y_k而言，其对应的分类超平面为$\langle \boldsymbol{w}_j - \boldsymbol{w}_k, \boldsymbol{x} \rangle + b_j - b_k = 0$，

因此式(42)考察样本$(\boldsymbol{x}_i, Y_i)$在各“相关-无关”标记配对情况下至分类超平面的距离，将其最小值定义为样本的分类间隔。基于此，学习系统在训练集$\mathcal{D}$上的分类间隔对应于：

$$\min_{(\boldsymbol{x}_i,Y_i)\in\mathcal{D}} \min_{(y_j,y_k)\in Y_i\times\bar{Y}_i} \frac{\langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k}{\|\boldsymbol{w}_j-\boldsymbol{w}_k\|} \tag{43}$$

理想情况下，假设上式定义的训练集分类间隔取值为正，即$\langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k > 0$ $(1\le i\le m,$ $(y_j,y_k)\in Y_i\times\bar{Y}_i)$成立。进一步地，通过对线性分类器$\mathbf{W}=\{(\boldsymbol{w}_j, b_j)\mid 1\le j\le q\}$的参数进行适当的缩放，从而使$\langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k \ge 1$ $(1\le i\le m,\ (y_j,y_k)\in Y_i\times\bar{Y}_i)$成立，且存在样本$(\boldsymbol{x}^*, Y^*)\in\mathcal{D}$及$(y_{j^*}, y_{k^*})\in Y^*\times\bar{Y}^*$使该式取等号。此时，最大化式(43)所示的训练集分类间隔可表述为如下优化问题：

$$\max_{\mathbf{W}} \min_{(\boldsymbol{x}_i,Y_i)\in\mathcal{D}} \min_{(y_j,y_k)\in Y_i\times\bar{Y}_i} \frac{1}{\|\boldsymbol{w}_j-\boldsymbol{w}_k\|^2} \tag{44}$$

$$\text{subject to}: \quad \langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k \ge 1 \quad (1\le i\le m, (y_j,y_k)\in Y_i\times\bar{Y}_i)$$

设训练样本足够充分，即对于所有类别标记y_j, y_k $(y_j\ne y_k)$，存在$(\boldsymbol{x}, Y)\in\mathcal{D}$使得$(y_j,y_k)\in Y\times\bar{Y}$。此时，上式的优化目标即对应于$\max_{\mathbf{W}}\min_{1\le j<k\le q}\frac{1}{\|\boldsymbol{w}_j-\boldsymbol{w}_k\|^2}$，相应的优化问题转化为：

$$\min_{\mathbf{W}} \max_{1\le j<k\le q} \|\boldsymbol{w}_j-\boldsymbol{w}_k\|^2 \tag{45}$$

$$\text{subject to}: \quad \langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k \ge 1 \quad (1\le i\le m, (y_j,y_k)\in Y_i\times\bar{Y}_i)$$

为了克服式(45)中的$\max$算子对优化造成的困难，Rank-SVM 算法将该算法子用求和算子加以近似，进一步地将上述优化问题转化为：

$$\min_{\mathbf{W}} \sum_{j=1}^{q} \|\boldsymbol{w}_j\|^2 \tag{46}$$

$$\text{subject to}: \quad \langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k \ge 1 \quad (1\le i\le m, (y_j,y_k)\in Y_i\times\bar{Y}_i)$$

为了反映真实情况下式(46)所示的约束无法完全满足的情况，可引入“松弛变量（slack variables）”改写算法对应的优化问题：

$$\min_{\{\mathbf{W},\Xi\}} \sum_{j=1}^{q} \|\boldsymbol{w}_j\|^2 + C\sum_{i=1}^{m} \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(y_j,y_k)\in Y_i\times\bar{Y}_i} \xi_{ijk} \tag{47}$$

$$\text{subject to}: \quad \langle \boldsymbol{w}_j-\boldsymbol{w}_k, \boldsymbol{x}_i\rangle + b_j - b_k \ge 1-\xi_{ijk}$$

$$\xi_{ijk}\ge 0 \quad (1\le i\le m, (y_j,y_k)\in Y_i\times\bar{Y}_i)$$

其中，$\Xi=\{\xi_{ijk}\mid 1\le i\le m,\ (y_j,y_k)\in Y_i\times\bar{Y}_i\}$为松弛变量集合。由上可见，式(47)中所示的目标

函数由两个求和项组成。其中，第一项对应于学习系统在训练集上的分类间隔（model complexity），第二项对应于学习系统在训练集上的经验误差（ranking loss），参数 C 用于平衡上述两项对目标函数的影响。

值得注意的是，式(47)对应于一个具有凸目标函数和线性约束条件的“二次规划（quadratic programming）”问题，但仅仅假设了线性模型用于样本分类。为了使得系统具有非线性分类能力，可以通过引入核技巧将式(47)转化成其“对偶形式（dual form）”求解，具体细节可参见文献[36]。

基于 2.1 节的符号表示，Rank-SVM 算法还采用了特殊的方式确定阈值函数$t(\cdot)$。具体来说，设$t(\boldsymbol{x}) = \langle \boldsymbol{w}^*, \boldsymbol{f}^*(\boldsymbol{x}) \rangle + b^*$为线性函数。其中，$\boldsymbol{f}^*(\boldsymbol{x}) = (f(\boldsymbol{x}, y_1), \cdots, f(\boldsymbol{x}, y_q))^{\mathrm{T}} \in \mathbb{R}^q$为 q 维属性向量，其分量对应于分类系统在各类别标记上的输出。相应地，$\boldsymbol{w}^* \in \mathbb{R}^q$为 q 维权值向量，$b^* \in \mathbb{R}$为偏置。给定训练集$\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le m\}$，Rank-SVM 使用线性最小二乘法求解相应参数：

$$\min_{\{\boldsymbol{w}^*, b^*\}} \quad \sum\nolimits_{i=1}^{m} \left(\langle \boldsymbol{w}^*, \boldsymbol{f}^*(\boldsymbol{x}_i) \rangle + b^* - s(\boldsymbol{x}_i) \right)^2 \tag{48}$$

$$\text{where} \quad s(\boldsymbol{x}_i) = \arg\min_{a \in \mathbb{R}} \left(\left| \{ y_j \mid y_j \in Y_i, f(\boldsymbol{x}_i, y_j) \le a \} \right| + \left| \{ y_k \mid y_k \in \bar{Y}_i, f(\boldsymbol{x}_i, y_k) \ge a \} \right| \right)$$

通常，$s(\boldsymbol{x}_i)$可能的取值对应于一个实数区间，算法取该区间的中值用以最小化式(48)。最终，基于式(47)与式(48)的解，则可得到所需的多标记分类器：

$$h(\boldsymbol{x}) = \{ y_j \mid \langle \boldsymbol{w}_j, \boldsymbol{x} \rangle + b_j > \langle \boldsymbol{w}^*, \boldsymbol{f}^*(\boldsymbol{x}) \rangle + b^*, \ 1 \le j \le q \} \tag{49}$$

3.3.3 LEAD

该算法的基本思想是基于“贝叶斯网络（Bayesian network）”对标记之间的相关性进行建模，并采用近似策略实现对贝叶斯网络的高效学习[22]。

基于 2.1 节的符号表示，设$\mathcal{X}$代表 d 维示例空间，$\mathcal{Y}$代表包含 q 个类别的标记空间，则学习系统的任务是构造多标记分类器$h : \mathcal{X} \to 2^{\mathcal{Y}}$。从贝叶斯学习的角度来看，该任务等价于对联合条件分布$p(\boldsymbol{y} \mid \boldsymbol{x})$进行建模，其中$\boldsymbol{x} \in \mathcal{X}$代表属性向量而$\boldsymbol{y} = (y_1, y_2, \cdots, y_q)^{\mathrm{T}} \in \{0, 1\}^q$为二值标记向量，其第 j 维用于指定属性向量$\boldsymbol{x}$具有（$y_j = 1$）或不具有（$y_j = 0$）第 j 个类别标记。

假设存在贝叶斯网络结构$\mathcal{G}$表征联合条件分布，则$p(\boldsymbol{y} \mid \boldsymbol{x})$可相应地分解为如下形式：

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod\nolimits_{j=1}^{q} p(y_j \mid \mathbf{pa}_j, \boldsymbol{x}) \tag{50}$$

其中$\mathbf{pa}_j$代表标记y_j在贝叶斯网络结构$\mathcal{G}$中对应的“父结点集合（parent labels）”。

值得注意的是，基于现有的贝叶斯网络结构学习方法，直接从训练集中学习所需的贝叶斯网络结构具有两方面的困难[37]：1) 结构学习算法需要处理具有混合类型的随机变量，其中属性向量$\boldsymbol{x}$通常为“连续变量（continuous variable）”而标记向量$\boldsymbol{y}$则为“离散变量（discrete variable）”；2) 属性

向量$\boldsymbol{x}$的维度往往过高，例如在文档分类问题中向量往往含有几千甚至几万个属性。由上可见，上述两种困难的存在其实都是由于属性向量$\boldsymbol{x}$参与贝叶斯网络结构学习而引起的。因此，LEAD 算法采用一种近似的方式首先消除属性向量对于标记空间的影响，然后再构造相应的贝叶斯网络来对标记之间的相关性进行建模。

基于上述考虑，LEAD 算法的具体步骤如下：

Step 1：针对标记空间$\mathcal{Y}$中的每一个类别y_j $(1 \le j \le q)$，基于训练集$\mathcal{D}_j$（式(16)）独立构造相应的二类分类器$g_j : \mathcal{X} \to \mathbb{R}$，并计算相应的预测误差$e_{ij} = [\![g_j(\boldsymbol{x}_i)]\!] - [\![y_j \in Y_i]\!]$ $(1 \le i \le m)$；

Step 2：基于预测误差e_{ij}学习得到贝叶斯网络结构$\mathcal{G}$；

Step 3：针对每一个类别y_j，将y_j的父结点集合$\mathbf{pa}_j$与属性集合$\boldsymbol{x}$组合作为输入属性，基于训练集$\mathcal{D}_j^*$构造新的二类分类器$g_j^* : \boldsymbol{x} \bigcup \mathbf{pa}_j \to \mathbb{R}$。其中，设$\mathbf{pa}_j(\boldsymbol{x})$代表$\boldsymbol{x}$的标记向量在父结点集合$\mathbf{pa}_j$上的取值，则$\mathcal{D}_j^* = \{((\boldsymbol{x}_i, \mathbf{pa}_j(\boldsymbol{x}_i)), \phi(Y_i, y_j)) \mid 1 \le i \le m\}$；

Step 4: 给定未见示例，基于贝叶斯网络结构$\mathcal{G}$所确定的“标记顺序（causal order of labels）”，使用分类器$g_j^*(\cdot)$以及属性集合$\boldsymbol{x} \bigcup \mathbf{pa}_j$迭代预测其具有的类别标记集合。

总的来看，LEAD 算法具有如下几个显著特点：1) 标记之间的相关性基于贝叶斯网做了显式且精炼的表示；2) 基于贝叶斯网络结构可以考察标记之间任意阶的相关性，其中相关性对应的阶数由每个结点具有的父结点个数决定；3) 算法的模型复杂度与标记空间中标记的个数成线性关系。

本小节分别介绍了三种具有代表性的“算法适应”类型的多标记学习算法，即一阶方法 ML-*k*NN[27]、二阶方法 Rank-SVM[11]以及高阶方法 LEAD[22]。除此之外，目前还存在其他一些基于“算法适应”的一阶[5][12][38]、二阶[6][10][17][39][40][41]以及高阶[35][42]多标记学习算法。限于篇幅，这里不再做一一介绍。

4. 结束语

本章主要介绍了多标记学习框架的定义、面临的主要问题、常用的评价指标、以及几种代表性学习算法。值得注意的是，在前述标准多标记学习框架的基础上，研究者们还对其研究课题进行了一系列拓展：

- “噪音/弱”标记：在某些情况下（如错误的人工标记），对象的标记信息存在噪音，从而导致隶属于对象的标记不一定为其“真实（valid）”标记[43][44]；另一方面，在某些情况下（如遗漏的人工标记），对象的标记信息存在缺失，从而导致不隶属于对象的标记有可能为其真实标记[45]。
- 未标记数据：在真实世界问题中，往往可以以较小的代价获得大量未标记数据。此外，在多

表 1　多标记学习的相关学术资源

资源类型	资源链接与描述
Tutorial	http://www.ecmlpkdd2009.net/program/tutorials/learning-from-multi-label-data/ (In conjunction with ECML/PKDD 2009)
Active Groups	http://lamda.nju.edu.cn/ (LAMDA Group at Nanjing University, China) http://mlkd.csd.auth.gr/ (MLKD Group at Aristotle University of Thessaloniki, Greece) http://www.uni-marburg.de/fb12/kebi/ (KEBI Lab at Philipps-Universität Marburg, Germany) http://www.ke.tu-darmstadt.de/ (KE Group at Technische Universität Darmstadt, Germany) http://rakaposhi.eas.asu.edu/ai/ (AI Lab at Arizona State University, USA) http://www.cse.msu.edu/~rongjin/ (ML Group at Michigan State University, USA) http://research.microsoft.com/en-us/groups/mcg/ (Media Computing Group at MSRA, China)
Software	http://mulan.sourceforge.net/index.html (The MULAN [63] open-source Java library for learning from multi-label data) http://cse.seu.edu.cn/people/zhangml/Resources.htm#codes_mll (Matlab codes for learning from multi-label data)
Data Sets	http://mulan.sourceforge.net/datasets.html (Data sets from MULAN project) http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html (Data sets from LIBSVM) http://meka.sourceforge.net/ (Data sets from sourceforge.net)
Online Bibliography	http://www.citeulike.org/group/7105/tag/multilabel (Bibliographic information on more than 100 multi-label literatures maintained at CiteULike.org)

标记学习问题中，由于每个对象可能具有多个标记，这将显著增加获取已标记数据的难度。因此，很有必要对基于未标记数据的多标记学习进行深入研究，如“主动（active）”多标记学习[46][47][48][49][50]、“半监督/直推（semi-supervised/transductive）”多标记学习[51][52]等。

● 维度约简：高维数据广泛存在于文档分类、生物信息学、多媒体应用等真实世界问题中，其维度约简对于提升多标记学习技术求解相关问题的能力具有十分重要的意义。目前，已经出现了一些基于“过滤（filter）”策略[16][53][54]、“封装（wrapper）”策略[55][56]以及“过滤-封装”混合策略[57]的多标记降维方法。

● 大规模数据：在多标记学习问题中，大规模数据包含两个层面的含义。一方面，当标记空间包含大量类别标记时，部分多标记学习算法尤其是二阶与高阶方法将失效[58][59][60][61]；另一方面，当训练集中包含大量样本时，部分多标记学习算法如一阶方法可在算法性能与算法效率之间保持较好的平衡[62]。

除了本章第 1 节提到的相关学术活动外（脚注 1, 2），表 1 还给出了与多标记学习相关的一些学术资源，以供读者参考。作为国际机器学习界一个新的研究热点，笔者相信，在今后几年甚至更长的时间内，多标记学习的研究工作将进一步深入发展，新兴研究成果将会不断涌现。

参 考 文 献

[1] Mitchell T M. *Machine Learning*, New York: McGraw-Hill, 1997.

[2] Tsoumakas G, Katakis I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 2007, 3(3): 1-13.

[3] Tsoumakas G, Zhang M-L, Zhou Z-H. Learning from multi-label data. *Tutorial at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'09)*, Bled, Slovenia, 2009. [http://www.ecmlpkdd2009.net/wp-content/uploads/2009/08/learning-from-multi-label-data.pdf]

[4] Schapire R E, Singer, Y. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 2000, 39(2/3): 135-168.

[5] McCallum A. Multi-label text classification with a mixture model trained by EM. In: *Working Notes of the AAAI'99 Workshop on Text Learning*, Orlando, FL, 1999.

[6] Ueda N, Saito K. Parametric mixture models for multi-label text. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing Systems 15 (NIPS'02)*, Cambridge, MA: MIT Press, 2003, 721-728.

[7] Gao S, Wu W, Lee C-H, Chua T-S. A MFoM learning approach to robust multiclass multilabel text categorization. In: *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, Banff, Canada, 2004, 329-336.

[8] Boutell M R, Luo J, Shen X, Brown C M. Learning multi-label scene classification. *Pattern Recognition*, 2004, 37(9): 1757-1771.

[9] Snoek C G M, Worring M, van Gemert J C, Geusebroek J M, Smeulders A W M. The challenge problem for automated detection of 101 semantic concepts in multimedia. In: *Proceedings of the 14th ACM International Conference on Multimedia (ACM Multimedia'06)*, Santa Barbara, CA, 2006, 421-430.

[10] Qi G-J, Hua X-S, Rui Y, Tang J, Mei T, Zhang H-J. Correlative multi-label video annotation. In: *Proceedings of the 15th ACM International Conference on Multimedia (ACM Multimedia'07)*, Augsburg, Germany, 2007, 17-26.

[11] Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Dietterich T G, Becker S, Ghahramani Z, eds. *Advances in Neural Information Processing Systems 14 (NIPS'01)*, Cambridge, MA: MIT Press, 2002, 681-687.

[12] Clare A, King R D. Knowledge discovery in multi-label phenotype data. In: De Raedt L, Siebes A, eds. *Lecture Notes in Computer Science 2168*, Berlin: Springer, 2001, 42-53.

[13] Barutcuoglu Z, Schapire R E, Troyanskaya O G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 2006, 22(7): 830-836.

[14] Tang L, Rajan S, Narayanan V K. Large scale multi-label classification via metalabeler. In: *Proceedings of the 19th International Conference on World Wide Web (WWW'09)*, Madrid, Spain, 2009, 211-220.

[15] Yang B, Sun J-T, Wang T, Chen Z. Effective multi-label active learning for text classification. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, 2009, 917-926.

[16] Yu K, Yu S, Tresp V. Multi-label informed latent semantic indexing. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, Salvador, Brazil, 2005, 258-265.

[17] Zhu S, Ji X, Xu W, Gong Y. Multi-labelled classification using maximum entropy method. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*,

Salvador, Brazil, 2005, 274-281.

[18] Gopal S, Yang Y. Multilabel classification with meta-level features. In: *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*, Geneva, Switzerland, 2010, 315-322.

[19] Song Y, Zhang L, Giles L C. A sparse Gaussian processes classification framework for fast tag suggestions. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*, Napa Valley, CA, 2008, 93-102.

[20] Ozonat K, Young D. Towards a universal marketplace over the web: Statistical multi-label classification of service provider forms with simulated annealing. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, 2009, 1295-1303.

[21] Hüllermeier E, Fürnkranz J, Cheng W, Brinker K. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 2008, 172(16-17): 1897-1916.

[22] Zhang M-L, Zhang K. Multi-label learning by exploiting label dependency. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*, Washington, D. C., 2010, 999-1007.

[23] Tsoumakas G, Vlahavas I. Random *k*-labelsets: An ensemble method for multilabel classification. In: Kok J N, Koronacki J, de Mantaras R L, Matwin S, Mladenič D, Skowron A, eds. *Lecture Notes in Artificial Intelligence 4701*, Berlin: Springer, 2007, 406-417.

[24] Salton G. Developments in automated text retrieval. *Science*, 1991, 253(5023): 974-980.

[25] Dembczyński K, Waegeman W, Cheng W, Hüllermeier E. Regret analysis for performance metrics in multi-label classification: The case of hamming and subset zero-one loss. In: Balcázar J, Bonchi F, Gionis A, Sebag M, eds. *Lecture Notes in Artificial Intelligence 6321*, Berlin: Springer, 2010, 280-295.

[26] Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K. Multilabel classification via calibrated label ranking. *Machine Learning*, 2008, 73(2): 133-153.

[27] Zhang M-L, Zhou Z-H. ML-*k*NN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007, 40(7): 2038-2048.

[28] Comité F D, Gilleron R, Tommasi M. Learning multi-label alternating decision tree from texts and data. In: Perner P, Rosenfeld A, eds. *Lecture Notes in Computer Science 2734*, Berlin: Springer, 2003, 35-49.

[29] Brinker K, Fürnkranz J, Hüllermeier E. A unified model for multilabel classification and ranking. In: *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI'06)*, Riva del Garda, Italy, 2006, 489-493.

[30] Godbole S, Sarawagi S. Discriminative methods for multi-labeled classification. In: Dai H, Srikant R, Zhang C, eds. *Lecture Notes in Artificial Intelligence 3056*, Berlin: Springer, 2004, 22-30.

[31] Ji S, Tang L, Yu S, Ye J. Extracting shared subspace for multi-label classification. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, Las Vegas, NV, 2008, 381-389.

[32] Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets. In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*, Pisa, Italy, 2008, 995-1000.

[33] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Buntine W, Grobelnik M, Shawe-Taylor J, eds. *Lecture Notes in Artificial Intelligence 5782*, Berlin: Springer, 2009, 254-269.

[34] Dembczyński K, Cheng W, Hüllermeier E. Bayes optimal multilabel classification via probabilistic classifier chains. In: *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, Haifa, Israel, 2010, 279-286.

[35] Cheng W, Hüllermeier E. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 2009, 76(2-3): 211-225.

[36] Elisseeff A, Weston J. Kernel methods for multi-labelled classification and categorical regression problems. *Technical Report*, BIOwulf Technologies, 2001.

[37] Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*, Cambridge, MA: MIT Press, 2009.

[38] Wang H, Ding C, Huang H. Multi-label classification: Inconsistency and class balanced *k*-nearest neighbor. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, GA, 2010, 1264-1466.

[39] Ghamrawi N, McCallum A. Collective multi-label classification. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany, 2005, 195-200.

[40] Zhang M-L, Zhou Z-H. Multilabel neural networks with application to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(10): 1338-1351.

[41] Brinker K, Hüllermeier E. Case-based multilabel ranking. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India, 2007, 702-707.

[42] Yan R, Tešić J, Smith J R. Model-shared subspace boosting for multi-label classification. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, San Jose, CA, 2007, 834-843.

[43] Jin R, Ghahramani Z. Learning with multiple labels. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing Systems 15 (NIPS'02)*, Cambridge, MA: MIT Press, 2003, 897-904.

[44] Ozonat K, Young D. Towards a universal marketplace over the web: Statistical multi-label classification of service provider forms with simulated annealing. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, 2009, 1295-1303.

[45] Sun Y-Y, Zhang Y, Zhou Z-H. Multi-label learning with weak label. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, GA, 2010, 593-598.

[46] Brinker K. On active learning in multi-label classification. In: *Proceedings of the 29th Annual Conference of the German Classification Society (GfKl'05)*, Magdeburg, Germany, 2005, 206-213.

[47] Qi G-J, Hua X-S, Rui Y, Tang J, Zhang H-J. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(10): 1880-1897.

[48] Yang B, Sun J-T, Wang T, Chen Z. Effective multi-label active learning for text categorization. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Paris, France, 2009, 917-925.

[49] Esuli A, Sebastiani F. Active learning strategies for multi-label text categorization. In: Boughanem M, Berrut C, Mothe J, Soule-Dupuy C, eds. *Lecture Notes in Computer Science 5478*, Berlin: Springer, 2009, 102-113.

[50] Singh M, Curran E, Cunningham P. Active learning for multi-label image annotation. *Technical Report UCD-CSI-2009-01*, School of Computer Science and Informatics, University of Dublin, Ireland, 2009.

[51] Liu Y, Jin R, Yang L. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In:

Proceedings of the 21st AAAI Conference on Artificial Intelligence (AAAI'06), Boston, MA, 2006, 421-426.

[52] Chen G, Song Y, Wang F, Zhang C. Semi-supervised multi-label learning by solving a Sylvester equation. In: *Proceedings of the 8th SIAM International Conference on Data Mining (SDM'08)*, Atlanta, GA, 2008, 410-419.

[53] Zhang Y, Zhou Z-H. Multi-label dimensionality reduction via dependency maximization. In: *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI'08)*, Chicago, IL, 2008, 1503-1505.

[54] Zhang Y, Zhou Z-H. Multi-label dimensionality reduction via dependency maximization. *ACM Transactions on Knowledge Discovery from Data*, 2010, 4(3): Article 14.

[55] Ji S, Ye J. Linear dimensionality reduction for multi-label classification. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, Pasadena, CA, 2009, 1077-1082.

[56] Qian B, Davidson I. Semi-supervised dimension reduction for multi-label classification. In: *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, Atlanta, GA, 2010, 569-574.

[57] Zhang M-L, Peña J M, Robles V. Feature selection for multi-label naive bayes classification. *Information Sciences*, 2009, 179(19): 3218-3229.

[58] Tsoumakas G, Katakis I, Vlahavas I. Effective and efficient multilabel classification in domains with large number of labels. In: *Working Notes of the ECML/PKDD'08 Workshop on Mining Multidimensional Data*, Antwerp, Belgium, 2008.

[59] Loza Mencía E, Fürnkranz J. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In: Daelemans W, Goethals B, Morik K, eds. *Lecture Notes in Artificial Intelligence 5212*, Berlin: Springer, 2008, 50-65.

[60] Hsu D, Kakade S, Langford J, Zhang T. Multi-label prediction via compressed sensing. In: Bengio Y, Schuurmans D, Lafferty J, Williams C K I, Culotta A, eds. *Advances in Neural Information Processing Systems 22 (NIPS'09)*, Cambridge, MA: MIT Press, 2009, 772-780.

[61] Zhang X, Yuan Q, Zhao S, Fan W, Zheng W, Wang Z. Multi-label classification without the multi-label cost. In: *Proceedings of the 10th SIAM International Conference on Data Mining (SDM'10)*, Columbus, OH, 2010, 778-789.

[62] Hariharan B, Zelnik-Manor L, Vishwanathan S V N, Varma M. Large scale max-margin multi-label classification with priors. In: *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, Haifa, Israel, 2010, 423-430.

[63] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: Maimon O, Rokach L, eds. *Data Mining and Knowledge Discovery Handbook*, Berlin: Springer, 2010, 667-686.