多示例学习与多标记学习的研究

张敏灵 周志华

1、研究背景

“机器学习”是研究怎样通过计算机模拟或实现人类学习活动的科学，是人工智能的核心研究领域之一。自上世纪八十年代起，经过二十多年的蓬勃发展，机器学习已成为计算机科学技术中最受关注的研究领域之一。在机器学习中，监督学习(supervised learning)是研究得最多、应用最广泛的一种学习框架。

在该学习框架下，每个真实世界的对象由一个包含若干属性的示例进行描述。与此同时，该示例对应于一个概念标记以表达其语义信息。学习系统通过对训练集中具有概念标记的训练例进行学习，以尽可能正确地预测未见对象的概念标记。在传统监督学习框架下，真实世界的对象与其描述及概念标记之间都是一一对应的关系。一般认为，这样的学习问题是没有歧义性(ambiguity)的[1]。

然而，歧义性对象在真实世界中却是广泛存在的。例如，对于图 1(a)所示的图像对象，从内容上看该对象包含了多种自然景物的描述信息，从概念上看该对象同时具有山、树木和湖泊等多个概念标记。再如，对于图 1(b)所示的网页对象，从内容上看该对象包含了多个段落的描述信息，从概念上看该对象同时具有交

(a)

(b)

图 1. 歧义性对象 (a)自然场景图像 (b)新闻网页

通、经济甚至文化等多个概念标记。因此，这些真实世界的对象无论从内容描述还是概念标记上都出现了歧义性。显然，基于传统的监督学习框架将很难用一个示例来完整地进行对象描述，且对象所对应的概念标记也不再是唯一的了。

这里，我们主要考察两种处理歧义性对象的学习框架：多示例学习(multi-instance learning)[2]以及多标记学习(multi-label learning)[3]。多示例学习从输入空间，即内容表示上来考察对象的歧义性；而多标记学习则是从输出空间，也就是概念标记上来考察对象的歧义性。本文接下来首先将从问题起源、研究现状以及我们的研究成果分别对这两种学习框架进行介绍。

此外，通过对歧义性学习问题的深入研究，我们还提出了多示例多标记学习(multi-instance multi-label learning)[4,5]这一新型机器学习框架。多示例多标记学从输入空间和输出空间两个方面同时考察对象的歧义性，相比于多示例学习或者多标记学习，可以更加自然且有效地处理歧义性对象。最后，本文将对进一步的研究工作进行展望。

2、多示例学习

2.1 问题起源

上世纪 90 年代中期，Dietterich 等人[2]对药物活性预测问题进行了研究。该问题的输入对象是一个分子，其输出是该分子与某个目标“绑定区域”耦合的紧密程度。对适于制造药物的分子来说，它的某个低能形状和期望的绑定区域将耦合得很紧密；而对不适于制造药物的分子来说，它和期望的绑定区域将耦合得不好。学习系统通过对已知适于或不适于制药的分子进行分析，以尽可能正确地预测某种新的分子是否适合制造这种药物。

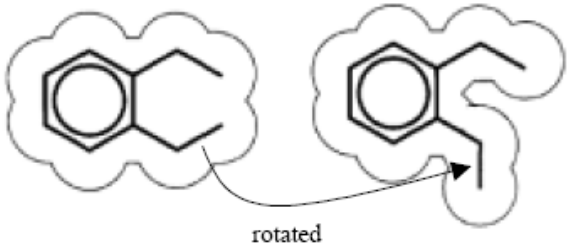

图 2.　分子的形状随着“内部键”的转动而变化[2]

该问题的困难之处在于，每一个分子有很多种可能的低能形状，只要该分子在其中一种形状下与期望的绑定区域紧密耦合，该分子就适于制造药物，如图 2 所示。生物化学专家只知道哪些分子适于制药，并不知道具体的哪一种形状起到了决定性作用，这就使得预测新的适于制药的分子成为一个非常困难的问题。

一种直观解决上述问题的方法是将正包中所有的示例看作正例，将反包中所有的示例看作反例，从而利用传统的监督学习方法求解。然而，Dietterich 等人[2]的实验结果表明，由于正包中大量存在的伪正例(false positive instances)而引入的噪音，上述方法很难有效地解决多示例学习问题。为此，他们将每个分子作为一个包，将分子的不同低能形状作为包中的示例，从而提出了多示例学习的概念。

在多示例学习中，每个训练包由多个示例组成，示例没有概念标记，但每个训练包有一个概念标记。如果包中至少有一个示例是正例，则该包被标记为正包；如果包中没有任何一个示例是正例，即所有示例都是反例，则该包被标记为反包。在此基础上，Dietterich 等人提出了三种“轴平行矩形(APR)”学习方法，并与常见的 C4.5 决策树、BP 神经网络学习方法进行了实验比较，发现基于多示例学习概念的 APR 学习方法可以较好地解决上述药物活性预测问题，而未考虑多示例学习自身特点的学习方法其效果很不理想。

多示例学习框架的提出在国际机器学习界引起了极大的反响，目前在该领域已经取得了大量的研究成果。本文接下来将简要介绍多示例学习的研究现状。

2.2 研究现状

Dietterich 等人使用 APR 方法解决药物活性预测问题的工作刚一公布，即得到了理论机器学习界的广泛关注。研究者们对 APR 在多示例学习框架下的 PAC 可学习性进行了研究，并获得了一系列的研究成果。研究表明，当示例的属性之间相互独立且包中的示例相互独立的情况下，APR 在多示例学习框架下是 PAC 可学习的。此外，如果包中的示例不满足相互独立性，那么在多示例学习框架下对 APR 进行 PAC 学习，与在传统的监督学习框架下对 DNF 公式进行 PAC 学习具有相同的难度，而后者则是一个 NP 完全问题。

除了多示例学习的理论研究之外，研究者们还设计了多种多示例学习算法。一种设计多示例学习算法的途径是将算法的注意焦点从对示例的区分转化为对

包的区分，而从传统的单示例监督学习算法转化而来。基于该思路，研究者们相继对各种监督学习算法进行了扩展，提出了多样性密度算法、多示例懒惰学习算法、多示例决策树、多示例神经网络、多示例核方法等等。如果我们将多示例学习中对象的表示形式从包转化为单示例，则可以方便地利用已有的单示例监督学习算法，从而为多示例学习算法的实现提供了第二种途径。现有的一些多示例学习算法，如 CCE、BARTMIP、MILES 等，都是沿着上述思路设计而来的。目前，多示例学习技术已经在基于内容的图像检索、场景分类、股票选择、界标匹配、计算机安全、Web 挖掘、计算机辅助诊断等诸多领域得到了广泛应用。

早期的多示例学习主要关注输出为离散值的多示例分类问题，但在药物活性预测等许多应用领域，如果能采用实值表示的输出，则更有助于问题的解决。后来，一些研究者开始关注输出为实值的多示例回归问题，相继提出了基于 EM 方法、*k* 近邻算法、多样性密度算法等的多示例回归算法。

多示例学习还引起了来自于归纳逻辑程序设计(inductive logic programming，简称 ILP)领域的研究者们的关注。机器学习常用的属性-值对表示方法实际上是一种命题知识表示，而 ILP 领域的研究使用的是一阶知识表示。相比于命题知识表示，一阶知识表示的表达能力要强的多，但要高效地对其进行学习却非常困难。De Raedt 认为，多示例表示为命题知识表示和一阶知识表示建立了联系，在这种表示下，既有助于利用一阶表示的表达能力又有助于发挥命题学习的高效性。

除了上述的标准多示例学习问题，研究者们还提出了若干广义多示例学习(generalized multi-instance learning)问题。在广义多示例学习问题中，包的概念标记往往不再取决一个底层概念，而会同时受到多个底层概念的影响。例如，在基于出现的多示例学习(presence-based multi-instance learning)问题中，一个包被标记为正包当且仅当对于概念集合中的每一个概念，包中均有一个示例与之对应。

有关多示例学习更加详细的研究内容，感兴趣的读者可进一步参见文献[6,7]。本文接下来将给出我们在多示例学习领域内取得的一些研究成果。

2.3 研究成果

我们主要从新型算法、算法设计思想、研究领域扩展这三个层面对多示例学习进行了研究。

（一）新型多示例学习算法

神经网络作为一种重要的机器学习方法，具有泛化能力强、预测速度快等优点。Dietterich 等人在提出多示例学习概念时即指出，设计基于神经网络的多示例学习算法是一个非常值得研究的课题。目前，已经相继出现了一些多示例版本的神经网络学习算法。然而，这些算法虽然取得了较好的实验结果，但其性能仍差于部分常用的多示例学习算法，如 MI Kernel、Iterated-discrim APR 等。因此，设计出泛化能力更强的神经网络的多示例版本是一个值得深入研究的问题。

为此，我们提出了一种基于径向基函数神经网络(radial basis function，简称 RBF)的多示例学习算法 RBF-MIP[8]。该算法对 RBF 神经网络的两层拓扑结构进行了改造以适应包的表示形式。具体来说，RBF-MIP 采用与 RBF 神经网络类似的两层拓扑结构，网络第一层中每个结点对应于一组训练包构成的簇，网络第二层的连接权则通过最小化网络的实际输出与期望输出之间的误差平方和优化得到。实验结果表明，其性能明显优于现有的多示例神经网络学习算法以及其他一些多示例学习算法。在此基础上，我们还将成功地将 RBF-MIP 算法用于基于内容的图像检索领域。

（二）多示例学习算法设计思想

一般来说，多示例学习问题的学习难度主要来自于其所采用的对象表示形式。在该学习框架下，每个对象采用一组属性向量的表示形式，即包的表示。已有的研究工作表明，传统的单示例监督学习算法可以通过将其注意焦点从对示例的区分转化为对包的区分，从而得到相应的多示例学习算法。实际上，大多数已有的多示例学习算法即是通过上述准则将传统的单示例学习算法进行改造以适应包的表示形式。反之，如果能设计出一种方法将包的表示形式合理地转化为单示例的表示形式，即可利用大量成熟的监督学习技术来处理多示例学习问题，为该问题的解决提供了一条新的途径。

基于上述思想，我们提出通过“表示转换”将多示例样本转化为单示例样本进行学习，这与目前将单示例算法改造为多示例算法的思路显著不同。相应的，我们给出了一种基于构造性聚类的集成学习方法 CCE (Constructive Clustering-based Ensemble)[9]。该方法首先通过一个聚类过程将对象所采用的包的

表示形式转化为单个示例的表示形式，从而将原始的多示例学习问题转化为传统的监督学习问题。此后，采用集成学习技术对转化后的监督学习问题进行学习。由于 CCE 通过聚类过程对对象的表示形式进行了转化，因此可以将该聚类过程看作一种特殊的构造性归纳机制。实验结果表明，CCE 算法在标准多示例学习以及广义多示例学习问题上均取得了较好的应用效果。

（三）多示例学习研究领域扩展

目前，多示例学习的研究主要集中于监督多示例学习问题（即多示例预测问题），而非监督多示例学习问题尚未引起研究者们的关注。然而，对于该问题的研究却具有十分重要的意义。首先，在许多情况下，获取包的概念标记往往比较困难或者代价较高。例如，虽然生物化学家可以较容易地设计出各种药物分子，并基于药物分子的多种低能形状将其表示成为一个包，但为了确定该分子的功能特性（即包的概念标记）却需要进行代价很高的生物化学实验。其次，考虑到非监督学习有助于发现数据集的内在结构信息。因此，即使对于包的概念标记已知的监督多示例学习问题而言，也可利用非监督多示例学习技术对数据集进行预处理，以期获得某些关于数据集的有用信息以便进一步的预测处理。

基于上述分析，我们对非监督多示例学习问题进行了研究并提出了一种多示例聚类算法 BAMIC (BAg-level Multi-Instance Clustering)[7]。该算法利用 Hausdorff 度量计算包之间的距离，并利用 *k*-Medoids 算法将原始的未标记的训练集划分为 *k* 个不相交的子集，每个子集对应于一组训练包构成的簇。此外，基于 BAMIC 算法的聚类结果，我们还给出了一种多示例预测算法 BARTMIP[7]。该算法通过将每个包转化为一个 *k* 维属性向量以实现包的表示形式转化，其中向量的第 *i* 维对应于包和第 *i* 簇“中心”之间的距离。实验结果表明，BAMIC 算法能够有效地发现多示例数据集的内在结构信息。与此同时，BARTMIP 算法在标准多示例学习、多示例回归以及广义多示例学习问题上均取得了很好的应用效果。

3、多标记学习

3.1 问题起源

多标记学习[3]最初起源于文档分类中所遇到的歧义性问题。例如，在文档分

类问题中，每篇文档可能同时隶属于多个预定义的主题，如“政府”与“健康”。实际上，除了文档分类问题，多标记学习问题还广泛存在于其他一些真实世界的问题当中。

举例来说，在功能基因组学问题中，每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”；在场景分类问题中，每幅场景图像可能同时包含了多种语义信息，如“海滩”与“城市”；在视频自动标注问题中，每个视频片断可能同时对应于多个语义类别，如“城市”与“建筑”等等。对于上述这些多标记学习问题，训练集中的每个示例均对应于一组概念标记，学习系统通过对多标记示例构成的训练集进行学习，以尽可能正确地预测训练集之外的示例的概念集合。

如果限定每个样本只对应于一个概念标记，那么传统的二类以及多类学习问题均可看作多标记学习问题的特例。然而另一方面，多标记学习问题的一般性使得解决该问题的难度大大增加。一种直观地解决多标记学习问题的方法是将其分解为多个独立的二类分类问题来求解，其中每个二类分类问题对应于一个可能的概念类。然而，由于该类方法没有考虑到每个样本所对应的概念标记之间的相关性，因此其泛化性能往往并不理想。例如，在文档文类问题中，如果已知一篇文档隶属于体育新闻类，则该文档同时隶属于休闲新闻类的可能性将大于其隶属于政治新闻类的可能性。再比如，如果已知一段视频或一幅图像隶属于“野生动物”类，则该视频或图像同时隶属于“草原”类的可能性将大于其隶属于“城市”类的可能性。因此，多标记学习问题的主要难点就在于如何充分利用各训练样本所含多个概念标记之间的相关性，从而有效地预测未知样本的概念标记集合。

目前，有关多标记学习已经取得了大量的研究成果。本文接下来将对多标记学习的研究现状进行简要介绍。

3.2 研究现状

目前，多标记学习已在文本分类、生物信息学、场景分类等许多领域得到了广泛应用。本节接下来根据多标记学习不同的应用领域，介绍该学习框架的研究现状。

如 3.1 节所述，多标记学习的研究起源于文档分类中遇到的歧义性问题。早

期，Schapire 和 Singer[3]为了将同一个文档归入多个类别，对经典的 AdaBoost 算法进行了扩展，提出了一种基于集成学习的文档分类系统 BoosTexter。在此之后，多标记学习逐渐引起了学者们的广泛关注。在 Bag-of-Words 的文档表示形式下，研究者们先后提出了多种“产生式概率模型”用于多标记文档分类。值得注意的是，这些基于贝叶斯理论的方法均需利用文档中出现的字的频率进行学习，因此仅适于解决文档分类这一特殊的多标记学习问题。

除此之外，目前还出现了许多由传统机器学习算法扩展而来的多标记文档分类算法。从技术路线来看，这些多标记学习算法主要采用了三种不同的设计策略。第一种策略是简单地将多标记学习问题分解为若干独立的二类分类问题求解，代表性算法有 ML-*k*NN、ML-SVM 等。第二种策略是考察标记之间的排序关系，将多标记学习问题转化为对标记的排序问题，代表性算法有 Rank-SVM、BP-MLL 等。最后一种是策略利用标记之间的相关性来提高学习系统的泛化性能，代表性算法有 CNMF、CLP 等。此外，还有些研究者提出了利用额外的类别层次信息或者未标记数据信息来提高多标记学习系统性能的方法。

除了文档分类问题，多标记学习还被应用于生物信息学领域。Clare 和 King 通过修改 C4.5 决策树中有关熵的定义来处理多标记数据，该方法的优点是决策树的输出可以转化为一组等价的符号规则从而可以与已知的生物知识进行比较。Elisseeff 和 Weston 通过定义一个称为“ranking loss”的代价函数和相应的间隔，提出了一种基于 SVM 的方法并在酵母基因功能分类问题上取得了较好的效果。Barutcuoglu 等人利用已有的基因功能分类系统所提供的结构信息，提出了一种贝叶斯学习框架来进行基因功能预测。

Boutell 等人通过将多标记学习问题转化为多个独立的二类学习问题，给出了一种基于多标记学习技术的场景分类方法。该方法在构造与某个概念类对应的二类分类器时，将所有包含该类的样本作为正例而将所有不包含该类的样本作为反例。他们还给出了多种根据各个二类分类器的输出来确定测试样本标记集的预测准则。此外，多标记学习还在计算机视觉、关联规则挖掘等领域中得到了成功应用。

有关多标记学习更加详细的研究内容，感兴趣的读者可进一步参见文献[10,11]。本文接下来将给出我们在多标记学习领域内取得的一些研究成果。

3.3 研究成果

目前已经出现了基于 SVM、决策树等传统机器学习算法的多标记学习算法。作为重要的机器学习技术，神经网络、*k* 近邻以及朴素贝叶斯在很多应用领域都取得了成功。因此，设计出基于这些机器学习技术的多标记学习算法，将有助于充分利用这些方面已有的研究成果来解决真实世界的多标记学习问题。

我们对传统的 BP 神经网络进行了扩展，提出了一种多标记神经网络学习算法 BP-MLL (BP for Multi-Label Learning)[12]。该算法通过对传统 BP 神经网络的全局误差函数进行改造，提出了一种新颖的误差函数以反映多标记学习问题的特性。简单地说，在对所有可能的概念标记排序时，隶属于示例的概念标记应位于不属于该示例的概念标记前列。实验结果表明，BP-MLL 算法在功能基因组学以及文档分类两种多标记学习问题上均取得了很好的应用效果。

我们还对传统的 *k* 近邻学习算法进行了扩展，提出了一种多标记懒惰学习算法 ML-*k*NN (Multi-Label *k*-Nearest Neighbor)[11]。对于给定的测试样本，该算法首先确定其在训练集中的 *k* 个近邻。然后，基于近邻样本的概念标记集合所蕴含的统计信息，ML-*k*NN 利用最大化后验概率准则确定测试样本的标记集合。在若干多标记学习问题上的应用表明，ML-*k*NN 算法的性能优于其他一些常用的多标记学习算法。

除了神经网络和 *k* 近邻之外，我们还对朴素贝叶斯方法进行了扩展，提出了一种多标记贝叶斯学习算法 MLNB (Multi-Label Naive Bayes)。我们将属性选择机制引入 MLNB 中以提高其算法性能。首先，MLNB 算法利用主成分分析技术去除数据集中的无关与冗余属性。然后，MLNB 算法利用遗传算法选取具有最优分类性能的属性子集。实验结果表明，MLNB 算法能够有效地处理数据集中存在的无关与冗余属性，获得较好的泛化性能。

总的来看，ML-*k*NN 的算法性能略优于 BP-MLL 以及 MLNB。然而另一方面，后两种学习算法却具有其自身的特点。BP-MLL 作为一种神经网络算法，与 ML-*k*NN 这种懒惰学习算法相比，其预测时间开销很小。因此，在要求进行快速预测的任务中，BP-MLL 是更好的选择；MLNB 内嵌了属性选择机制，算法性能受无关属性以及冗余属性的影响较小。因此，在对具有大量无关或冗余属性的数

据集进行学习时，MLNB 可能是更好的选择。

4、多示例多标记学习

如第 1 节所述，在传统的监督学习框架下，每个对象由一个示例表示且对应于一个概念标记。虽然监督学习框架在许多领域内得到了广泛应用，但对于真实世界中的某些学习问题，采用单示例单标记的对象表示形式往往是不合适的。

例如，在图像分类问题中，一幅图像通常包含多个区域其中每个区域可以由一个示例进行表示，与此同时该图像可能同时隶属于“山”与“树”等多个概念标记；在文本分类问题中，每个文档通常包含多个段落其中每个段落可以由一个示例进行表示，而该文档在从不同的角度进行考察时可能同时隶属于“科幻小说”、“儒勒·凡尔纳作品”甚至“旅游书籍”类；在 Web 挖掘中，网页中的每个链接可以由一个示例进行表示而网页本身则可能隶属于“新闻页面”、“体育页面”以及“足球页面”等。

考虑到真实世界的对象在内容描述和概念标记上均可能出现歧义性，我们提出了一种新型的处理歧义性对象的机器学习框架，即多示例多标记学习[4,5]。在

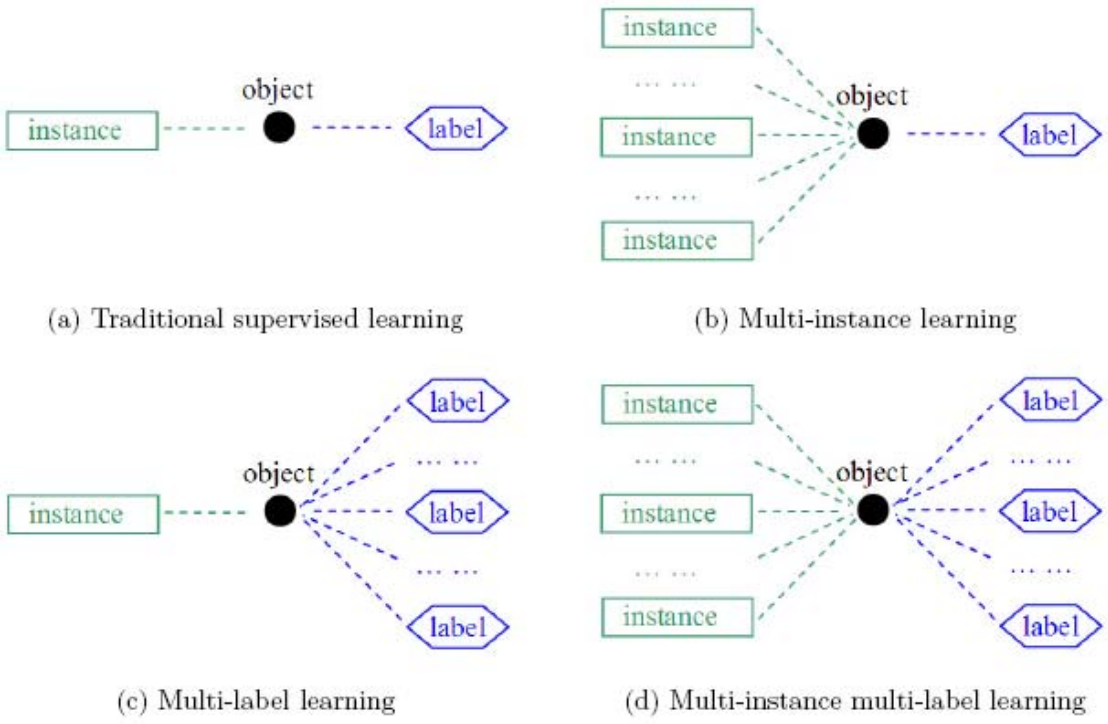

图 3.　四种不同的学习框架[4]

该学习框架下，每个对象由多个示例表示且同时隶属于多个概念标记。对于歧义性对象而言，采用多示例多标记的表示形式显得更加自然，可以方便地同时刻画对象的输入歧义性和输出歧义性。图 3 对多示例多标记学习与传统监督学习、多示例学习以及多标记学习之间的关系进行了比较。

如图所示，很显然传统监督学习可以看作多示例学习或者多标记学习的“特例”，而传统监督学习、多示例学习以及多标记学习均可看作多示例多标记学习的特例。因此，一种直观地解决多示例多标记学习问题的策略是以多示例学习或者多标记学习为桥梁，将其退化为传统的监督学习问题进行求解。我们沿着这一思路设计了 MIMLBoost 和 MIMLSVM 算法，可以有效地对多示例多标记学习问题求解。如果能设计出不通过退化的方法，那么多示例多标记学习可望发挥更大的作用。

已有的研究结果表明，对歧义性对象而言，基于多示例多标记学习框架进行建模将有助于相应学习问题的解决。然而，在不少应用中，我们只能获得已由他人进行了特征提取、已将一个对象表示为一个特征向量的数据，而不能利用多示例多标记表示直接对原始对象进行建模。但即使在这种情况下，我们的研究结果表明，多示例多标记学习框架仍然能发挥重要的作用。

事实上，对于采用单示例多标记表示形式的对象，此时该对象多个概念标记所蕴含的多样性信息仅仅内嵌于单一的示例中。因此，如果能将对象单一示例的表示形式合适地转化为包（一组示例）的表示形式，使得包中的每个示例均能从特定方面清晰地反映对象所包含的某种信息，那么将有助于学习问题的解决。

基于上述思想，我们提出了一种基于示例区分(instance differentiation)的学习算法 INSDIF。该算法的执行主要包括两个阶段，在第一阶段，INSDIF 将每个样本转化为包的表示形式从而在输入空间中显式地描述对象歧义性。简单地说，对于每一个可能的类别 c_l 构造一个原型向量 $\mathbf{v}_l$，该向量为具有类别 c_l 的所有训练样本对应的均值向量。基于此，将每个样本 $\mathbf{t}$ 转化为包的表示形式，包中的每个示例对应于向量 $\mathbf{t}-\mathbf{v}_l$，即该样本与类别 c_l 的原型向量之间的差值。在算法的第一阶段完成后，初始的多标记学习问题即转化为了一个多示例多标记学习问题。在算法的第二阶段，INSDIF 利用多示例多标记学习器对转化后的数据集进行学习。目前我们使用了一种两层的分类策略来实现上述目的。实验表明，基于示例

区分技术将多标记学习问题转化为多示例多标记学习问题求解，可以获得更优的学习性能。

5、结束语

本文主要考察了两种歧义性对象学习框架，即多示例学习和多标记学习。我们从问题起源、研究现状以及所做工作三个方面分别对这两个框架进行了介绍。此外，本文还介绍了多示例多标记学习这一新型机器学习框架。传统的监督学习、多示例学习以及多标记学习均可看作多示例多标记学习的特例。在该学习框架下，歧义性对象所蕴涵的输入歧义性和输出歧义性可以很自然地表达出来，从而有利于对歧义性对象的进一步处理。

作为一种新型机器学习框架，有关多示例多标记学习还有许多问题值得深入研究。例如，如果能够从理论上清楚地分析出多示例多标记学习技术在解决实际问题时为什么有效、在什么情况下有效，那么对于多示例多标记学习研究的发展将具有重要的指导意义。

现有的学习算法基于退化的策略将多示例多标记学习问题转化为传统监督学习问题求解。然而，是否能够通过考察示例与标记之间的关系，设计出直接针对多示例多标记样本的学习算法将是一个值得研究的问题。实际上，我们在这方面已经进行了初步的尝试，提出了多示例多标记间隔的概念并设计了相应的算法 M^3MIML[14]。实验结果表明，在充分考察示例与标记之间的关系后，M^3MIML 算法可以获得比基于退化策略的多示例多标记学习算法更优的性能。

目前，多示例多标记学习已成功地应用于场景分类以及文档分类问题[4,5]。将该学习技术用于更多真实世界问题的求解，例如基于内容的图像检索、生物信息学等，也是一个值得研究的方向。

参考文献

[1] Maron O. Learning from ambiguity. PhD dissertation, Department of Electrical and Computer Science, MIT, Cambridge, MA, Jun. 1998.

[2] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence, 1997, 89(1-2): 31-71.

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

[4] Zhou Z-H, Zhang M-L. Multi-instance multi-label learning with application to scene classification. In: Schölkopf B, Platt J, Hofmann T, eds. Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, 2007, 1609-1616.

[5] Zhou Z-H, Zhang M-L, Huang S-J, Li Y-F. MIML: A framework for learning with ambiguous objects. *CORR abs/0808.3231*, 2008.

[6] 周志华. 多示例学习. 见: 刘大有 主编, 知识科学中的基本问题研究, 北京: 清华大学出版社, 2006, 322-336.

[7] Zhang M-L, Zhou Z-H. Multi-instance clustering with applications to multi-instance prediction. Applied Intelligence, in press since Jan. 2008, DOI 10.1007/s10489-007-0111-x.

[8] Zhang M-L, Zhou Z-H. Adapting RBF neural networks to multi-instance learning. Neural Processing Letters, 2006, 23(1): 1-26.

[9] Zhou Z-H, Zhang M-L. Solving multi-instance problems with classifier ensemble based on constructive clustering. Knowledge and Information Systems, 2007, 11(2): 155-170.

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

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

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

[13] Zhang M-L, Zhou Z-H. Multi-label learning by instance differentiation. In: Proceedings of the 22nd Conference on Artificial Intelligence, Vancouver, Canada, 2007, 669-674.

[14] Zhang M-L, Zhou Z-H. M^3MIML: A maximum margin method for multi-instance multi-label learning. In: Proceedings of the 8th IEEE International Conference on Data Mining, Pisa, Italy, 2008, 688-697.

作者简介

张敏灵，男，1979 年 4 月生。河海大学计算机及信息工程学院讲师。2007 年于南京大学计算机科学与技术系获博士学位。中国计算机学会会员，江苏省计算机学会人工智能专委会委员。主要研究兴趣为机器学习与数据挖掘。

周志华，男，1973 年 11 月生。南京大学计算机科学与技术系教授，博士生导师，

教育部长江学者特聘教授。2000 年于南京大学计算机科学与技术系获博士学位。中国计算机学会高级会员，人工智能与模式识别专业委员会副主任。主要研究领域为人工智能，机器学习，数据挖掘等。