DOI: 10.15960/j.cnki.issn.1007-6093.2018.02.001

多示例学习问题研究进展综述*

田英杰[1,2] 胥栋宽[1] 张春华[3,†]

摘要 多示例学习是一种特殊的机器学习问题，近年来得到了广泛的关注和研究，许多不同类型的多示例学习算法被提出，用以处理各个领域中的实际问题. 针对多示例学习的算法研究和应用进行了较为详细的综述, 介绍了多示例学习的各种背景假设, 从基于示例水平、包水平、嵌入空间三个方面对多示例学习的常见算法进行了描述, 并给出了多示例学习的算法拓展和若干领域的主要应用.

关键词 多示例学习, 分类问题, 包, 支持向量机, 深度学习

中图分类号 O234

2010 数学分类号 91E40, 68T05

A review of multi-instance learning research*

TIAN Yingjie[1,2] XU Dongkuan[1,2] ZHANG Chunhua[3,†]

Abstract Multi-instance learning is a special kind of machine learning problem, has received extensive attention and been researched on in recent years. Many different types of multi-instance learning algorithms have been proposed to deal with practical problems in various fields. This paper reviews the algorithm research and application of multi-instance learning in detail, introduces various background assumptions, and introduces multi-instance learning from three aspects: instance level, bag level, and embedded space. Finally we provide the algorithm extensions and major applications in several areas.

Keywords multi-instance learning, classification problem, bag, support vector machine, deep learning

Chinese Library Classification O234

2010 Mathematics Subject Classification 91E40, 68T05

收稿日期：2017-09-30

* 基金项目：国家自然科学基金(Nos. 71731009, 61472390, 71331005, 91546201, 11771038), 北京自然科学基金(No.1162005)

1. 中国科学院虚拟经济与数据科学研究中心, 大数据挖掘与知识管理重点实验室, 北京 100190; Research Center on Fictitious Economy and Data Science, the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

2. 中国科学院大学经济与管理学院，北京 100190; School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

3. 中国人民大学信息学院, 北京 100872; School of Information, Renmin University of China, Beijing 100872, China

† 通信作者 E-mail: zhangchunhua@ruc.edu.cn

0 引言

多示例学习起源于药物分子的活性检测[1], 其目标是通过分析一组已知的药物分子, 预测一个新的药物分子是否具有特定的药物活性. 药物分子活性检测的主要难点在于, 一个药物分子有很多不同的低能量状态对应的形状, 但是我们只知道某个药物分子是否具有活性, 而不知道具体是它的哪一个或者几个低能量状态具有相应的活性. 一个直观解决此类问题的方式是借助于传统的有监督学习方法, 把一个具有活性的药物分子所有低能量状态的形态都当做具有药物活性, 把不具有活性的药物分子所有低能量状态的形态都当做不具有药物活性. 遗憾的是, 由于一个药物分子的低能量状态对应的形态中只有一部分甚至小部分才真正具有药物活性, 这样就导致训练样本标签的不准确性. 因此, 直接借用传统的有监督学习来处理该问题不能取得较好的结果.

Dietterich等人[1]将这类新的学习问题描述为多示例学习问题. 与标准的分类问题不同, 多示例学习问题给定的训练集为

$$\{(\mathcal{X}_1, \mathcal{Y}_1), \cdots, (\mathcal{X}_l, \mathcal{Y}_l)\}, \tag{0.1}$$

其中输入$\mathcal{X}_i$是空间$\mathbb{R}^n$上有限个点组成的集合, $\mathcal{X}_i = \{x_{i1}, \cdots, x_{il_i}\}, x_{ij} \in \mathbb{R}^n, j = 1, \cdots, l_i; i = 1, \cdots, l$, 输出$\mathcal{Y}_i$是对$\mathcal{X}_i$的类别标号, $\mathcal{Y}_i \in \{-1, 1\}, i = 1, \cdots, l$, 其目的是据此找出一个规则, 用以推断任一 $\mathbb{R}^n$ 上有限个点组成的集合 $\tilde{\mathcal{X}} = \{\tilde{x}_1, \cdots, \tilde{x}_m\}$ 对应的类别标号 $\mathcal{Y}$. 这里称空间 $\mathbb{R}^n$上的点为示例, 而称有限个示例组成的集合为包. 因此训练集(0.1) 中的输入$\mathcal{X}_i$ 都是包, 分别称对应的 $\mathcal{Y}$ 值为 1 和 -1 的包为正包和负包. 可见上述多示例分类问题就是在已知若干个正包和若干个负包的情况下, 推断一个新的包是正包还是负包. 图1描述了空间 $\mathbb{R}^2$上的多示例两类分类问题的一个训练集, 这里每一个圈起的部分是一个包, "+" 和 "o" 都表示包中的示例; 包含用 "+" 表示的示例的包是正包, 包含用 "o" 表示的示例的包是负包. 需要解决的问题是, 如何推断由平面上有限个点组成的集合是正包还是负包.

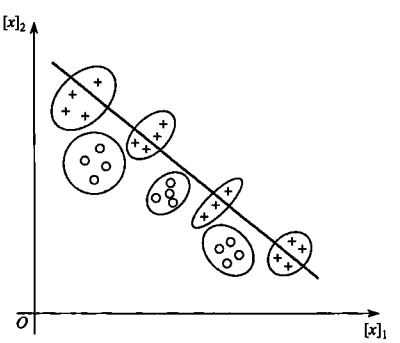


图1 多示例两类分类问题示意图

具体到药物分子的活性检测问题而言, 由于每种低能形态可以用空间 $\mathbb{R}^n$ 上的一个向量表示, 所以低能形态可以对应上述问题中的示例, 而药物分子则对应上述问题中的包. 由于药物实验只能检测出药物分子是否有活性, 而无法检测出它的哪种低能形态是否有活性, 所以只能对包进行标号. 然而, 药物分子是否有活性取决于它是否有一种低能形态是否有活性, 所以在研究包的类别时, 需要对示例引进类别标号, 把该标号与包的类别标号联系起来.

定义 0.1(包与示例的类别关系) 一个包是正包的充分必要条件是它的示例中至少有一个是正的, 一个包是负包的充分必要条件是它的所有示例都是负的.

由定义0.1, 上述涉及多示例的分类问题便归结为如下通常所说的多示例分类问题.

定义 0.2(多示例分类问题) 给定训练集(0.1), 其中, 当$\mathcal{Y}_i=1$ 时, $(\mathcal{X}_i,\mathcal{Y}_i)$ 的含义是正包$\mathcal{X}_i=\{x_{i1},\cdots,x_{il_i}\}$ 中至少有一个示例x_{ij} 属于正类; 当$\mathcal{Y}_i=-1$ 时, $(\mathcal{X}_i,\mathcal{Y}_i)$ 的含义是负包$\mathcal{X}_i=\{x_{i1},\cdots,x_{il_i}\}$ 中所有的示例x_{ij} 都属于负类. 据此寻找空间 $\mathbb{R}^n$ 上的一个实值函数 $g(x)$, 以使用决策函数

$$f(x)=\operatorname{sgn}(g(x)) \tag{0.2}$$

推断空间 $\mathbb{R}^n$ 上的任一示例 x 对应的标号y.

显然, 上述推断示例类别标号的决策函数 $f(x)$ 可以直接用来推断一个包 $\tilde{\mathcal{X}}=\{\tilde{x}_1,\cdots,\tilde{x}_m\}$ 的类别. 事实上, 只有该包中所有示例 $\tilde{x}_1,\cdots,\tilde{x}_p$都被推断为负类时, 该包才被推断为负类; 否则被推断为正类, 即 $\tilde{\mathcal{X}}$ 的类别标号 $\tilde{\mathcal{Y}}$ 应取为

$$\tilde{\mathcal{Y}}=\operatorname{sgn}\Big(\max_{i=1,\cdots,m} f(\tilde{x}_i)\Big). \tag{0.3}$$

在过去的这些年里, 有很多不同类型的多示例学习算法被提出, 用以处理在各个领域中的多示例学习问题[2,3]. 本文针对多示例学习进行较为详细的综述, 全文结构如下:第1节介绍多示例学习的各种背景假设, 为多示例学习建立灵活的基础框架; 第2节从基于示例水平、包水平、嵌入空间三个方面对多示例学习的常见算法进行了详细的描述; 第3节介绍多示例学习的算法拓展, 为多示例学习框架移植到其他学习领域提供了一定的借鉴; 第4节给出了多示例学习在若干领域的主要应用; 最后给出总结与展望.

1 基本假设

如上所述, 多示例学习区别于传统有监督学习. 在多示例学习中, 训练对象分为两个层次—包和示例; 而在传统有监督学习中, 训练对象与特征向量具有一对一的关系; 在多示例学习中, 只有包的准确标签信息, 示例的标签信息不完全; 而在传统有监督学习中, 每个对象都有一个标签信息, 用来指导算法模型的学习. 因此我们称多示例学习是一种弱监督学习框架. 对于现有的多示例学习算法而言, 都存在针对示例标签和包标签之间的一种关系的假设, 该假设是具体多示例学习方法的基础, 比如, 我们上节给的定义0.2 就是建立在最为常见的多示例学习假设——标准多示例学习假设(Standard multi-instance assumption, SMI)的基础上. SMI 假设每个正包中至少含有一个正示例, 而负包中的示例全是负示例. 但是, 在实际生活中还存在很多情况不满足SMI 成立的条件. 而且基于具有针对性的背景假设建立应用模型越来越引起人们的兴趣. 我们主要介绍四类常见的假设[4], 同时分析不同假设的使用背景.

基于存在的假设 基于存在的假设是多示例学习里面应用范围最广的假设. 在这类假设下, 如果一个包里面包含一个或者多个正示例, 那这个包就是正包. 标准的多示例学习假设SMI就属于这类假设. 这类假设的主要特点是, 包的特征信息是由局部信息决定, 而且局部信息只要存在, 包信息就会改变. 主要应用领域有药物分子活性检测[1]、图像分类[5]、文本分类[6]. 药物分子(包)中只要存在一种同分异构体(示例)具有药物活性, 那该药物分子就是属于有效的药物分子; 在常规的图片特征提取框架下, 只要图片(包)的某块(示例)含有特定目标的信息特征, 则该图像就可以被分类; 同理, 一篇文本(包)可以被视为由

有很多章节(示例)构成, 只要其所包含的一个章节包含有特定信息内容, 则该文本就能够被分类.

基于阈值的假设 在这类假设下, 如果一个包里面存在至少一定数量的正示例, 那这个包就是正包. 相比较于基于存在的假设, 基于阈值的假设更加宽泛, 对于一个正包而言, 所要求最少的正示例的个数是有最低要求的. 这是因为在某些场景中, 一个包的标签信息是由至少一定量的局部信息所决定的[4]. 比如在一个特定的图像分类问题中, 要求图片(包)中至少包含有5个"人"(正示例所对应的内容信息)的图片才是我们感兴趣的图片(正包).

基于计数的假设 在这类假设下, 如果一个包里面包含有一定数量范围的正示例, 那这个包就是正包. 相比较于前两种假设, 基于计数的假设设置了一个正包包含正示例个数的上限[4]. 基于计数的假设也有直接的应用场景, 比如在西方国家的领导人选举投票中, 假设有很多的地区(包)都进行领导人选举, 一种分析策略是关注比较摇摆的地区, 因为如果能够提前识别和分析这样的地区, 对于某一方的选举具有很大的帮助. 这样, 我们就可以把比较摇摆的地区视为正包, 它的特点是支持某一方的选取人(正示例)票数在一定的中间范围的区间内.

基于群体的假设 在这类假设下, 一个包的标签信息是由这个包里面的所有示例决定的, 而且, 各个示例对于包的标签信息的贡献度是一样的. 基于群体的假设是由概率论衍生的一个多示例学习基本假设. 在这个假设下, 一个包可以被看成是在示例空间上的一个概率分布, 每一个观测到的示例是从这个概率分布中随机采样得到的[7]. 相较于SMI假设只考虑包中部分的示例信息, 基于群体的假设会考虑包中的所有示例的信息.

除了以上四种基本的假设外, 还存在一些特定的假设. 在这些特定的多示例学习假设中, 有些假设本质上还是以上四种假设之一, 只不过是具有特定的模型结构, 比如文献[8]提出的软包假设. 总体而言, 多示例学习发展至今已经拓展了很多不同的假设模型, 基于特定的应用领域和处理问题而设定具体的多示例学习框架, 越来越受到相关领域研究者的重视, 尤其是图像、视频研究领域.

2 常见算法分析

多示例学习算法可以分为基于示例水平空间的多示例算法、基于包水平空间的多示例算法和基于嵌入空间的多示例算法.

2.1 基于示例水平空间的多示例算法

在这类算法中, 假设具有区别度的信息是存在于示例水平. 因此, 这类算法的核心思想是:首先, 训练一个示例水平的分类器, 使之能够区分来自正负包中的示例. 然后, 对于一个新来的包, 我们用训练好的示例水平的分类器对该包中的每个示例进行判断, 最后, 集成该包所有示例的判断值而得到对该包的判断值. 在基于标准的多示例背景假设下的多示例学习算法基本思想如图2所示, 其中左图中正包的示例来自于Class 1(被视为正示例)和Class 3, 负包的示例来自非Class 1. 分类器是在示例水平空间训练而得的, 从而可以对每个示例进行判断. 常见的基于示例水平空间的多示例算法有APR[1]、mi-SVM[6]、SMILE[9]、SVR-SVM[10]、Clustering MIL[11].

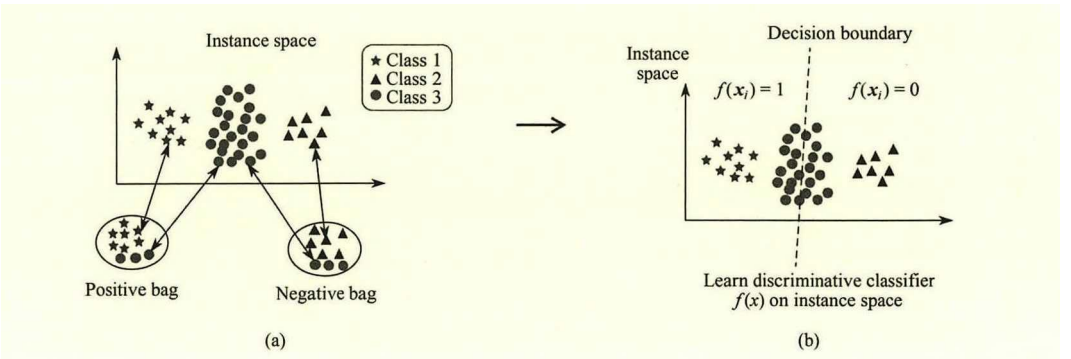


图2 基于示例水平空间的多示例算法示意图[3]

APR算法 源自于对药物分子的活性检测. 基于射线的分子结构表示方法, 药物分子可以被表示为一个长向量. 如图3(a)所示, 不规则封闭曲线代表了一个药物分子的同分异构体, 计算由原点出发的八条射线与曲线的交点到原点之间的距离就可以得到一个八维的向量$(x_1, x_2, \cdots, x_8)^{\top}$.

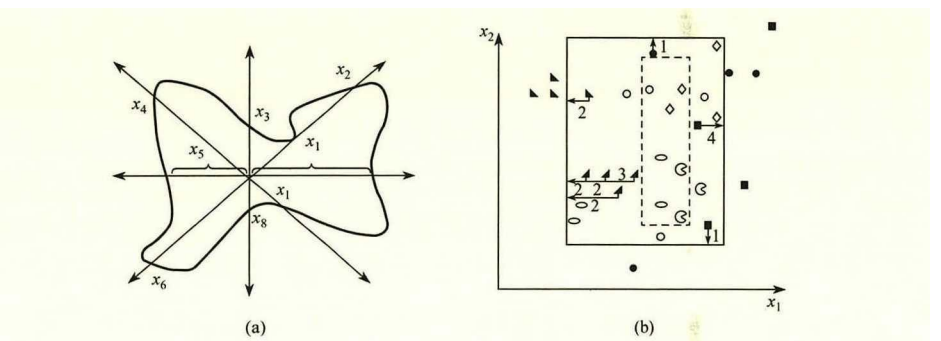


图3 APR算法示意图[1]

APR算法 有三种子算法: 1) GFS elim-count APR算法. 如图3(b)所示, 其中颜色和形状相同的点代表属于同一个包的示例, 黑色代表属于负包, 白色代表属于正包, 白色点旁的数字代表排除该示例所需要付出的代价. 该算法首先构造平行于轴的实线矩阵, 然后通过贪心算法排除属于负包的示例以不断缩小矩形的大小范围, 最终形成虚线的矩形; 2) GFS kde APR 算法. 其考虑矩形覆盖涉及到的属于正包的示例数, 在不断缩小矩形大小的时候, 尽量少的排除那些剩余示例较少的正包中的示例; 3) Iterated-discrim APR算法. 首先找出能够至少覆盖每个正包中一个示例的最小矩形, 然后挑选出最具有区别能力的一组属性, 接着通过核密度估计不断扩展矩形的边界.

mi-SVM算法 首先把所有正包的示例视为正示例, 基于这些标记的"正示例"和所有负包中的负示例训练一个标准的SVM分类器; 然后利用该SVM分类器重新标记训练集中的所有正包, 如果一个正包的所有示例都被判定为负, 则将该包中具有最大决策函数值的示例标记为正; 基于这些被标记为正的示例和所有负包中的示例重新训练支持向量机, 不断进行训练和标记, 直到所有训练示例的类别标签不再发生变化.

SMILE算法 基本流程如图4所示: 第一步从每个正包中选择一个示例组成初始正候选示例; 第二步分别计算每个示例与正负类的相似度; 第三步基于定义的数据模型, 用similarity-based SVM(SSVM)分类器进行训练; 第四步利用一个启发式的算法再次选择正候选示例并更新SSVM 分类器. 在测试阶段, 当一个包中被判断出存在一个正示例,

则该包就被判断为正示例.

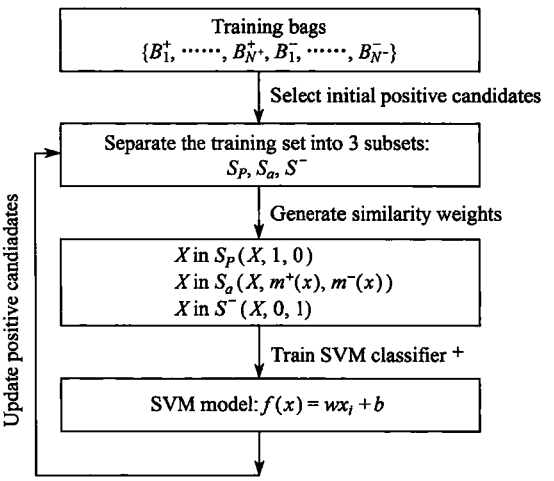


图4 SMILE算法流程图[9]

SVR-SVM算法 把标准多示例学习的非凸假设(当一个包中至少存在一个正示例, 则该包为正包; 当一个包中所有示例都为负示例, 则该包为负包) 重新定义为一个针对每个示例属于正负类别的似然比的凸约束, 进而把原始的多示例学习问题转化为一个针对示例的似然比函数和似然比函数值的凸的联合估计. 该算法采用了支持向量回归机去估计这个似然比, 取得了较好的实验效果. **Clustering MIL**通过对所有正包里面的示例进行聚类, 得到一些球形的区域, 并把某特定区域称为"概念"区域, 然后把落入概念区域的示例视为正示例.

2.2 基于包水平空间的多示例算法

在这类算法中, 包被视为一个整体来处理. 相较于基于示例水平空间的算法是在示例特征空间水平进行捕捉局部的信息, 基于包水平空间的多示例算法是在包水平空间执行具有区别度的学习过程. 这类算法的基本思想是: 首先定义一个度量包之间距离的函数, 然后把该距离函数嵌入标准的基于距离的分类器中, 比如SVM. 图5 所示为基于包水平空间的多示例算法基本思想的示意图, 其中图5(a) 代表训练的过程, 图5(b)代表测试的过程.

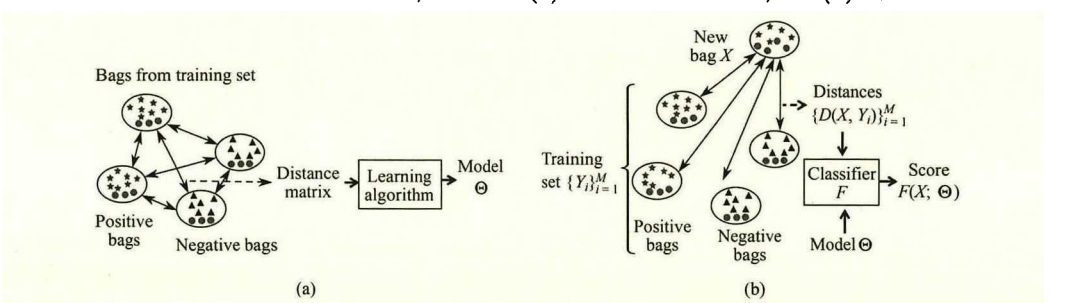


图5 基于包水平空间的多示例算法示意图[3]

基于包空间水平的多示例算法直接对包进行操作, 当进行包推理的时候, 相较于基于示例空间水平的多示例算法, 该类算法能够考虑更多的信息. 如图6所示, 假设正包是指包含两类示例的包, 负包是指只包含其中一类示例的包. 在这种情况下, 基于示例水平的算

法所训练出来的分类器只能区分单个的示例, 这样的分类器是不能区分正包和负包. 而基于包空间水平的分类器则能够直接区分不同包之间的差异, 从而能够处理这类数据情况.

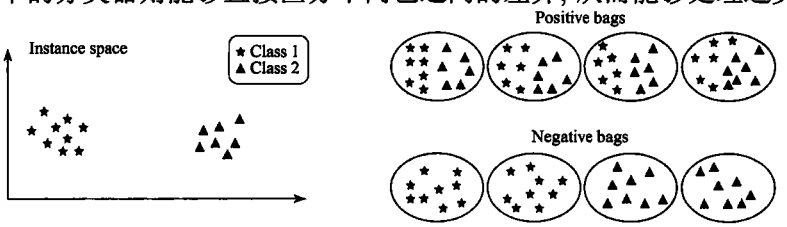


图6 对于基于包水平空间的多示例算法有效的、但基于示例水平的多示例算法无效的示例和包组成情况[3]

常见的基于包水平空间的多示例算法有MI-SVM[6]、 Bayesian-kNN[12]、 Citation-kNN[12]、MI-Kernel[13]、MIForest[14]、MInD[15].

MI-SVM算法 首先把每个包中的示例进行相加得到该包的基于示例水平的单特征向量表示; 然后利用所有包的单特征向量训练一个标准的SVM分类器; 接着利用该分类器去评估正包中的每个示例, 并且利用对应决策函数值最大的示例代替该包的单特征向量, 进而重新训练SVM分类器; 不断重复训练、转换, 直到训练集中的所有包的单特征向量都不再发生变化.

Bayesian-kNN和Citation-kNN是基于k最近邻的基本思想、采用minimum Hausdorff 距离来计算两个包之间的距离的多示例学习算法. 前者不仅考虑k近邻包的标签信息, 而且考虑k近邻包的先验概率, 从而预测新包的标签信息. 后者借用了科学文献中的"引用"的概念, 在预测新包的标签信息时, 不仅考虑k近邻包的标签信息, 还考虑将该包视为近邻包时的包标签信息. 这两个算法需要计算包之间的距离矩阵, 存储开销和计算时间成本比较大.

MI-Kernel算法 把正包里面的所有示例视作正示例, 把每个包里的示例进行规则化求和, 然后把该求和的结果来代表对应的包. 具体而言, 该方法采用集合核函数把包的特征投影到高维空间. 该核函数结合了SVM分类器可以简单有效的实现包的分类, 并且取得了较好的分类效果.

$$k(X,X')=\frac{k(X')}{f_{\text{norm}}(X)f_{\text{norm}}(X')}$$

MIForest算法 把随机森林拓展到了多示例学习算法中. 该算法的出发点是, 利用随机森林所具有的准确度高、速度快、易并行、能够处理多类别数据的特点. 为了找到示例的潜在类别标签, 该算法把示例标签视为一个概率分布上的随机变量, 通过迭代寻找能够使目标函数最小化的分布而消除示例标签的不确定性. 又由于该总体目标函数是非凸的, 在训练阶段, 该算法基于确定性退火算法提出了一个高效的求解算法. 值得注意的是, 该算法可以拓展到on-line 的学习问题.

MInD算法 直接采用一个包与其他在训练集中的包的区别作为该包的特征向量表示, 同时用不同的方式, 例如mean、EMD、Hausdorff 等定义包之间的区别, 讨论并比较了不同的差异定义对不同数据集的表现. 该算法具有计算成本不高, 同时和其他常见多示例算法相比, 表现出一定的竞争力.

2.3 基于嵌入空间的多示例算法

基于包水平空间的多示例算法和基于嵌入空间的多示例算法基本思想类似, 都是抽取有关包的全局的信息, 只不过, 在基于嵌入空间的多示例算法中, 这种信息抽取的方式

是一种隐形的, 比如, 定义一个距离映射函数D或者核函数K. 不一样的函数定义对不同类别的信息有不一样的侧重, 同时对模型算法的最终表现也有很大的影响.

这类算法的基本框架是: 每个包都会被映射为一个单一的特征向量, 该特征向量可以描述和该包相关的整体信息; 这样, 原始的包空间被映射为了一个向量化的嵌入空间, 在这个空间中进行分类器的训练, 从而原始的多示例问题就被转化为了一个标准的有监督学习问题. 图7所示为该基本思想的示意图, 其中图7(a)代表训练的过程, 图7(b)代表测试的过程.

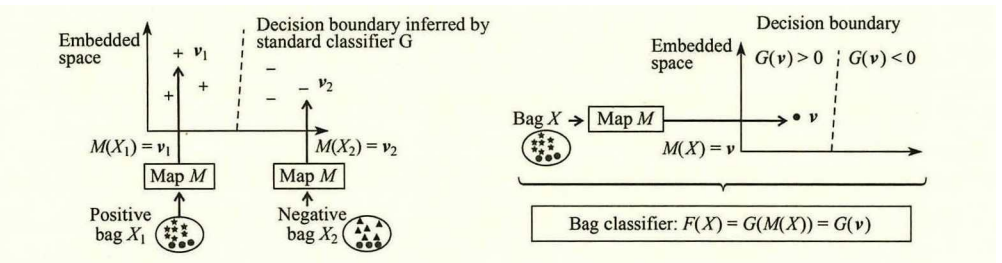


图7 基于嵌入空间的多示例学习算法示意图[3]

该类常见的算法可分为两种, 不基于词汇的学习算法和基于词汇的学习算法. 前者的代表算法是有Simple MI[16], 后者主要包括DD-SVM[17]、MILES[18]、MILD[19]、MILIS[20]、RSIS[21].

Simple MI算法 简单但具有一定效果, 核心思想是计算每个包里面示例在各个特征维度的平均值, 也就是对所有示例求取平均, 然后用这个计算得到的平均特征向量表示对应的包.

DD-SVM算法 基本思想: 基于多样性密度函数进行示例选择, 将训练集中多样性密度局部最大的示例作为原型, 然后基于找到的示例原型集对包进行映射, 接着用所有包的特征向量带入标准的高斯SVM中进行训练; 在测试阶段, 首先, 基于示例原型集, 将测试集中的包映射为特征向量, 然后带入训练好的分类器中. 其中, 多样性密度值取值0和1之间, 描述了来自不同正训练包中的示例的共现性. 如果一个示例和所有正包中的示例很近, 又与负包中的示例很远, 则该示例的多样性密度值会接近1. 如图8中所示, A代表一个多样性高密度区, B 仅代表一个高密度区. 寻找示例原型集是采用了启发式搜索策略, 即从训练集中的正包中的示例出发搜索具有最大多样性密度的示例, 以不同示例为起点进行多次的迭代产生示例原型.

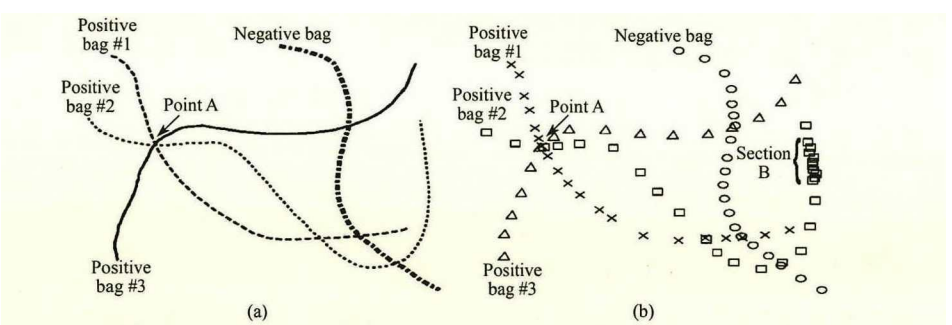


图8 DD-SVM算法示意图[17]

MILES算法 流程：首先基于训练集中的所有示例, 对每个训练包进行特征映射, 从而, 每个包就由一个新的相似性特征向量表示; 然后用所有训练集中的相似性特征向量学习一个标准的线性SVM, 由于对权重向量采用了1 范式的约束, 因此训练得到的权重w 会有某些元素为0, 也就实现了特征选择的效果; 更新特征向量并重新训练SVM 分类器. 值得注意的是, 由于包的每个特征对应一个示例, 因而挑选出来的特征对应被挑选的示例. 在训练阶段, 每个测试集中的包首先被映射到由所有示例原型形成的嵌入空间中, 获得对应的相似性特征向量, 最后将该特征向量输入训练好的SVM分类器中, 以获得该包的类别标签.

MILD算法 是基于一种类条件概率模型进行示例选择, 将每个正训练包中分类能力最强的示例作为原型, 然后基于这些示例原型对包进行映射, 从而得到每个包的特征向量. 挑选示例原型的基本思想是, 基于定义的决策函数的经验风险表达式(该表达式所定义的经验风险度量了在特定阈值下决策函数对训练包的分类准确率), 可以计算特定示例对所有训练包的分类能力, 因此, 我们从训练集中的每个正包中, 将经验风险值最大的示例挑出作为原型. 该算法的基本出发点是, 一个负示例到训练集中正负包的距离和一个正示例到训练集中正负包的距离不一致. 得到包的特征向量之后, 可以把这些向量带入标准的分类器中进行训练, 从而预测新包的标签信息.

MILIS算法 从训练集的每个包中选择一个示例原型, 然后通过一种迭代优化机制反复更新示例原型以及训练一个线性SVM分类器, 指导算法收敛. 具体而言, 该算法, 基于高斯核的核密度估计子

$$f(x)=\frac{1}{Z\sum_i m_i}\sum_{y_i=-1}\sum_{j=1}^{m_i}\exp(-\beta||x-x_{ij}||_2).$$

对正包选择一个具有最小似然值的示例(最有可能为正的示例), 对负包选择一个具有最大似然值的示例(最有可能为负的示例)作为示例原型. 然后基于示例原型对包进行特征映射, 进而带入标准的线性SVM 分类器中, 得到特征权重w. 接着固定w, 优化包的特征向量, 如此不断迭代, 直到所有示例原型不再发生变化. 测试阶段, 首先基于最终选择的示例原型对新包进行特征映射, 然后带入已经训练好的SVM分类器, 实现新包的类别标签预测.

RSIS算法 流程是, 首先在随机生成的特征子空间, 通过聚类对所有的示例计算Positivity Score, Positivity Score 值越大说明该示例越有可能是正示例; 然后为了生成一组多样性的分类器, 每个分类器会在不同的训练集子集上训练, 而这些训练子集包含了从每个正包中选择Positivity Score值最大的一个示例和所有负包的示例组成训练集. 该算法具有不需要考虑学习问题的背景假设以及能够处理低witness rate的特点. Positivity Score 计算如图9所示:

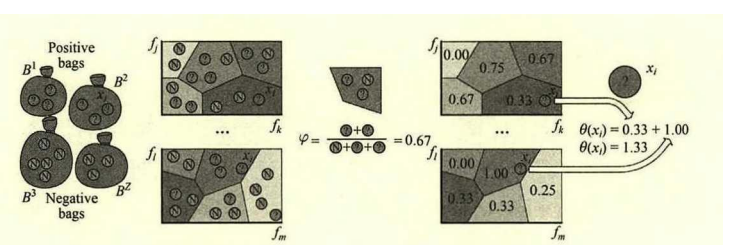


图9 RSIS的Positivity Score计算方法示意图[21]

3 算法拓展

3.1 深度学习

深度学习是近年来十分火热的研究技术, 在图像识别与分类、音频信息处理、自然语言处理等方面均具有十分优异的表现[22,23]. 现代信息社会中呈几何增长的大数据信息为深度学习框架提供了充分的训练资源, 从而保障了算法训练的有效性; 不断发展、增强的计算能力为深度学习的发展提供了有效的实现方式. 不少多示例学习方法充分利用了深度学习的优势, 表现出了更好学习效果.

图像与文本标注[23] 如图10所示, 该算法是第一个结合多示例学习与深度学习的算法, 算法思想朴实但具有借鉴意义: 依靠于多示例的学习框架, 利用深度学习方法, 分别对文本和图像信息进行分割处理, 最后对图像分类以及标签标注.

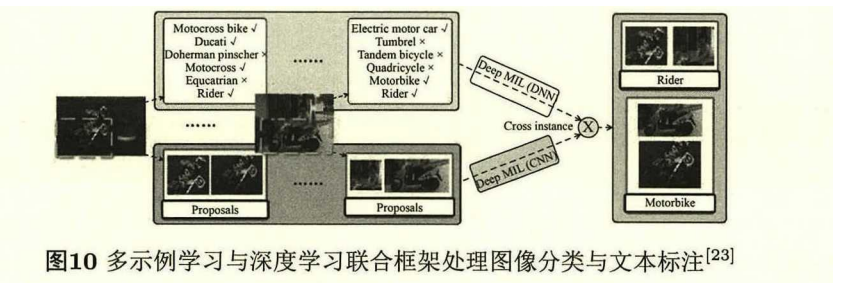


图10 多示例学习与深度学习联合框架处理图像分类与文本标注[23]

视频动作的检测[24] 视频是由很多帧组成的, 而具有目标动作信息内容的关键帧存在少数的卷中. 基于多示例学习框架, 我们可以把视频当做一个包, 视频所包含的卷被视为该视频的示例, 具有目标信息内容的卷视为正示例. 这样, 视频动作的检测问题就可以转化为对关键卷的挖掘与检测问题了. 该算法的框架如图11所示.

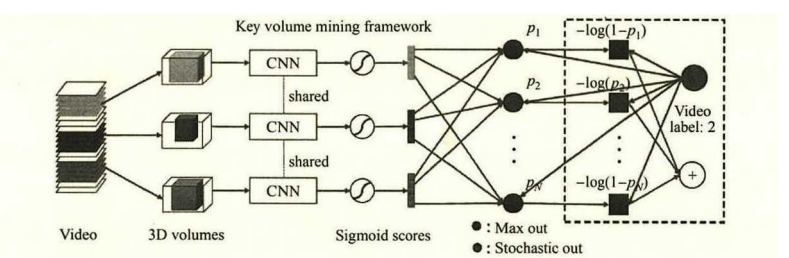


图11 多示例学习与深度学习联合框架进行视频动作检测[24]

3.2 隐马尔可夫模型

隐马尔可夫模型是一个统计模型, 它可以被用来描述一个含有隐含未知参数的马尔可夫过程, 从而更加深刻地描述目标对象的发展变化规律. 其难点是从可观察的参数中确定该过程的隐含参数, 然后利用这些参数来作进一步的分析. 文献[25]提出了一个基于自回归隐马尔可夫模型和多示例学习框架的活动识别算法模型. 我们可以通过信息接收器捕捉不同动作的信息量, 如图12所示.

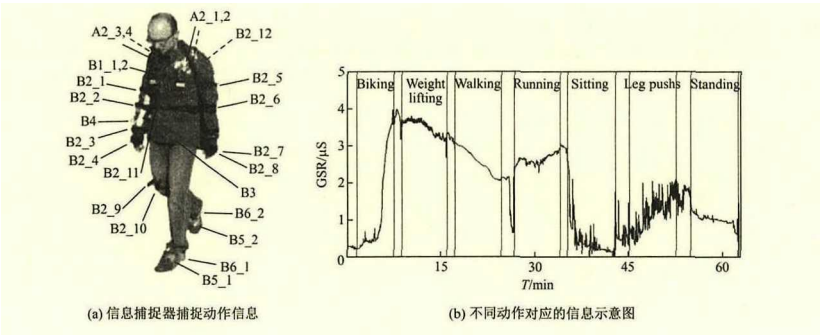


图12 个体不同动作的信息捕捉示意图[25]

该算法把一个大的时序区间内容的所有活动的信息当做一个包, 不同活动所对应的小时序区间的信息当做示例, 包含有特定活动内容的包被视为正包. 这样, 我们就可以用多示例的学习框架处理活动识别问题. 该算法进一步假设, 不同活动之间的变化是有潜在的变量所决定的, 活动之间的变化存在一定规律, 为了识别出潜在的变量和这种变化规律, 该算法基于隐马尔可夫模型和自回归模型, 对活动识别的多示例框架进行了拓展, 取得了很好的效果. 模型如图13所示[25], 其中, X为能够之间观察到的变化对象(示例), Z 为潜在变量, I为示例标签, Y为包标签, 剩余为模型参数.

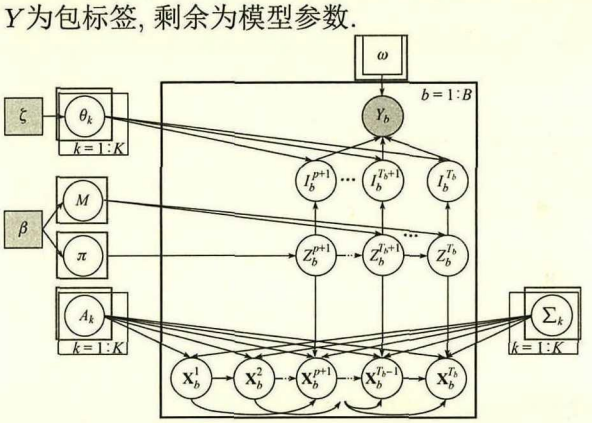


图13 多示例学习与深度学习联合框架进行视频动作检测[25]

3.3 局部敏感哈希学习

局部敏感哈希(locally sensitive hashing, LSH)是一种有效的信息检索技术[26]. 如下图所示, 和一般哈希技术相比, 局部敏感哈希能够有效地对相似的对象进行匹配—相似的对象在基于LSH的哈希映射空间中保持了相似性.

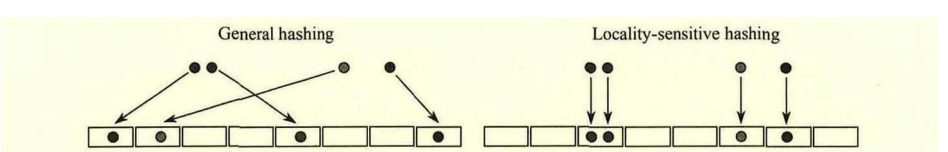


图14 局部敏感哈希技术与一般哈希技术的区别

文献[27]基于LSH提出了一个新颖的多示例学习方法FAMER. 该方法通过LSH, 把一个包中的示例映射到特定的空间, 进而把包转化成了对应映射空间中的直方图(描述了该

包中的示例在该映射空间中的分布情况), 最后基于包对应的、特有的特征向量, 设计了新的kernel. FAMER算法框架如图15所示, 其中, 圆点代表了示例, 图中所示的包含三个示例; 该包通过一组三个LSH函数转变为了01010100的特征向量, 最后连接通过很多组这样的LSH 函数得到的特征向量用来表示对应的包.

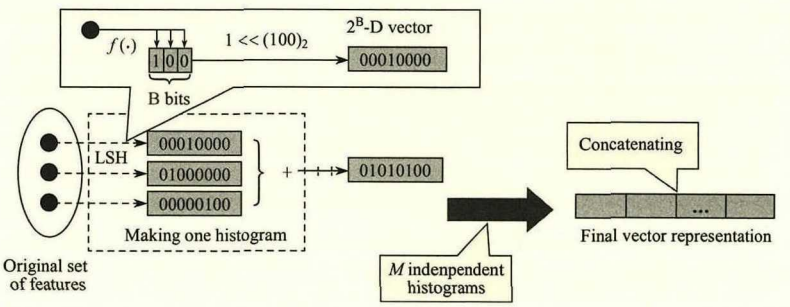


图15 FAMER算法框架示意图[27]

3.4 多示例多标记

现实生活中的对象往往并不只存在唯一的一个标签信息, 如图16所示, 我们既可以认为图16(a)具有“大象”的类别标签, 也可以认为它具有“草地”、“狮子”等标签; 同样, 我们也可以认为图16(b)可以具有“体育明星”的类别标签, 也可以认为它具有“娱乐”、“新闻”等标签. 结合多标记学习[28], Zhihua Zhou等人提出了多示例多标记学习[29], 该学习算法相较于传统监督学习、多示例学习和多标记学习最主要的特点是, 一个对象既含有多个示例, 也带有多个类别标记, 如图17所示.

(a) 例子1 (b) 例子2

图16 具有多标签的图像实例[29]

多示例多标记学习具有以下的特点: 传统监督学习问题和多示例学习问题可以视为MIML的特例; 对真实世界的学习问题求解中, 好的表示具有十分重要的意义. 使用MIML来对多义性对象进行表示, 有助于明示示例与类别标记之间的联系, 从而有助于学习任务的解决; 与简单地对合适标记进行预测相比, 了解一个对象为什么具有某个类别标记可能在某些场合具有更重要的意义, 而MIML 为此提供了一种可能性, 如图18所示.

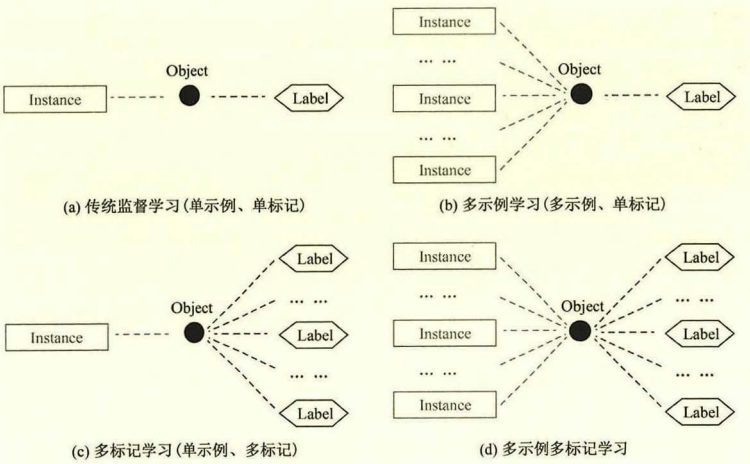


图17 传统监督学习、多示例学习、多标记学习、多示例多标记学习的区别与联系[29]

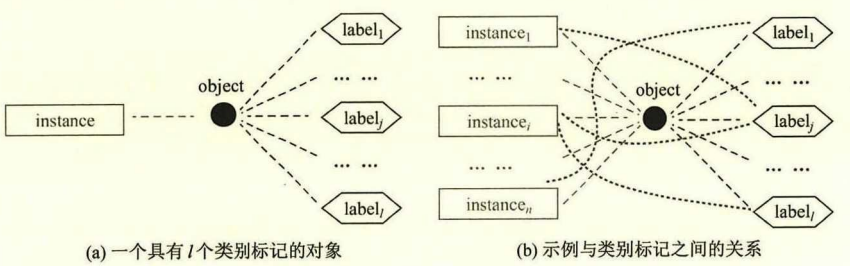


图18 多示例多标签学习的数据结构与标签[29]

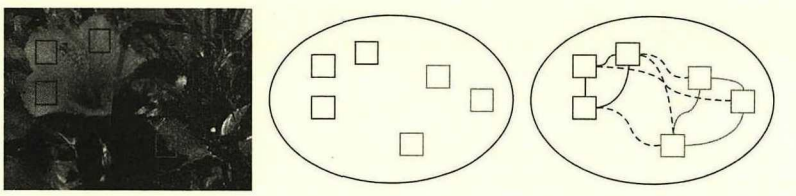
4 主要应用

由于多示例学习中训练样本的层次性表示结构, 多示例学习可以更好地反应现实中很多问题的逻辑结构, 在对“粗标签”对象的学习问题上, 具有特有的优势. 多示例学习已经得到了很广泛的应用, 主要包括药物分子活性检测、基于内容的图像检索、文本分类、目标识别、医疗图像辅助识别等.

4.1 基于内容的图像检索

基于内容的图像检索是多示例学习的一个重要应用领域[5]. 传统的基于内容的图像检索方法往往是基于一些底层的特征, 比如像素值、图像形状等, 对图像进行特征提取, 然后带入分类器进行训练.

多示例学习为这类研究领域提供了一种新的解决思维: 基于一幅图像的局部内容信息进行信息匹配和挖掘, 而非整幅图像的内容. 基本框架思想如图19(a)和(b)所示: 首先利用图像分割技术, 把一幅幅的图像分割为不断的子块, 具有“花”信息内容的子块被视为正示例; 然后对每个子块进行特征提取. 这样, 基于内容的图像检索就转化为了一个多示例学习问题: 每个子块被视为一个示例, 每张图像被视为一个包, 对图像的检索, 就转化为了对包的检索. 我们还可以基于这样的学习框架探究示例之间的关系, 以达到更好的学习表现, 如图19(c)所示.



(a) 基于多示例对图片分割 (b) 建立包和示例的模型 (c) 考虑示例之间的关系

图19 基于多示例学习的图片分类学习框架[30]

4.2 文本分类

随着信息技术的发展, 文本信息挖掘得到了广泛的关注和研究, 文本分类是文本信息挖掘中的一个重要研究领域. 传统的文本分类技术主要是对单个文本进行词频特征提取, 然后把文本转化为一个向量表示而进一步处理.

多示例学习也为文本分类提供了一个新的解决问题的框架: 不同的文章可以看做是不同的相互交叉的文本章节的集合, 每篇文章可以被视为一个包, 每篇文章里的章节可以被视为属于该文章的示例[6]. 基本学习框架如图20所示, 图20左侧是原始文章的示意图, 图20右侧中被矩形圈住的是属于该文章的章节. 进而, 我们可以根据我们的目标主题确定正类别, 从而把文本分类的问题, 转化为了一个多示例学习问题. 值得注意的是, 某篇文章和我们的目标主题相关(属于正类)并不意味该篇文章的所有章节都和我们的目标主题相关.

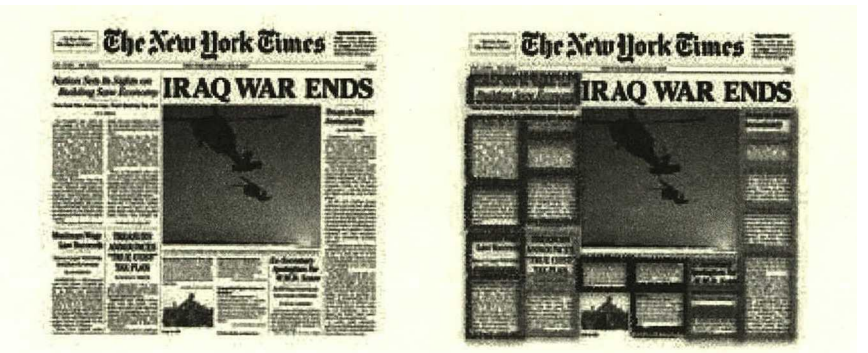


图20 基于多示例学习的文本分类学习框架[6]

4.3 目标识别

目标识别具有十分广泛的应用, 比如视频内容检测、人脸识别、指纹识别. 传统的目标识别方法通常是采用有监督的学习方法, 对目标物体进行标记、带入分类器训练、实现预测. 但传统方法存在很多不足, 比如, 我们获取训练样本的成本比较高, 需要人工去标注, 而且标注也存在困难和不准确性.

多示例学习同时为目标识别提供了一种解决问题的办法[30]. 采用多示例学习方法处理目标识别的出发点是, 我们不需要精确知道目标内容的具体标记. 比如, 对于检测某图片中是否含有特定的内容, 我们只需要标记该图片是我们“感兴趣的”, 而不用具体去标注图片中哪部分是对应于我们感兴趣的内容, 这样可以大大降低人工标注的难度和成本. 如下图21所示, 标准的目标检测方法需要对图像中的人进行较为精确的人工标注, 这样一方面存在较大的标注成本, 同时不一定能够准确的标注出图像中人的区域; 右图则示意了MIL学习处理目标识别问题的时候, 仅仅需要对整幅图像进行标注即可.

4.4 医疗图像辅助识别

在最近过去的十年里, 信息技术深刻的影响着现代医疗技术的发展, 计算机辅助诊断就是其中很重要的发展领域. 计算机辅助诊断是指医生借用计算机科学技术, 基于病人生成的病理检测数据, 对病人进行各种症状的诊断. 但是伴随着计算机辅助诊断也存在不少问题, 一个比较突出的问题是对数据标注的准确度. 以CT片为例, 因为医生不是计算机专家, 所以医生很难十分准确地标注出一块具有病变的区域, 这样导致标注的区域中既可能存在病变的组织区域, 也可能存在正常组织对应的区域, 最终会导致在我们的训练集中存在较大的噪音, 影响学习的准确性.

多示例学习医疗图像辅助诊断中存在的这类问题提供了一个有效的解决方法: 对于病变区域, 基于多示例学习的方法把整个病变区域视为正包, 通过图像分割技术对病变区域进行分割, 每个被分割的小区域就被视为该正包中的一个示例; 对于非病变区域, 同样也进行图像分割, 整个非病变区域视为负包(也就是我们不感兴趣的区域), 把被分割的子块视为负示例. 这样就可以用多示例学习的方法来解决这个问题. 如图22所示是一个针对医疗图像辅助识别的基本学习框架.

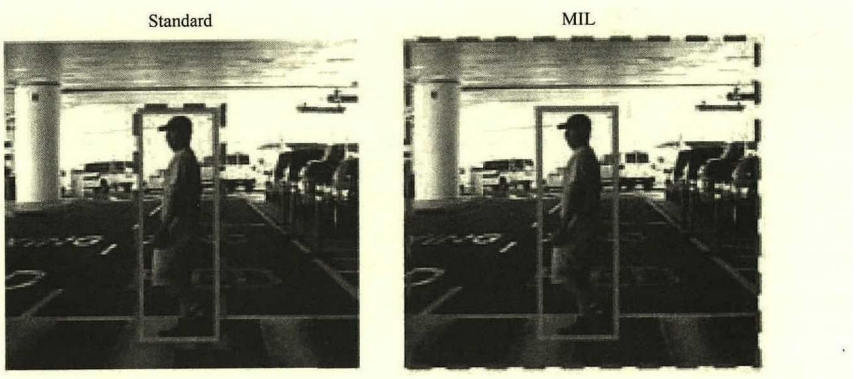


图21 基于多示例学习的目标识别学习框架[31]

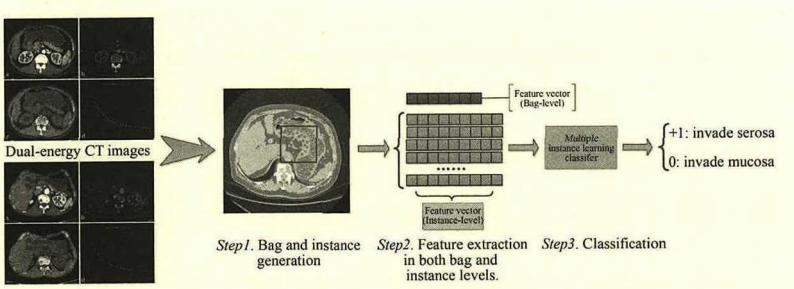


图22 基于多示例学习的医疗图像辅助识别学习框架[8]

5 结论

本文系统梳理了多示例学习的研究特点. 首先介绍了多示例学习的各种背景假设, 为多示例学习建立了灵活且强大的基础框架, 接着从基于示例水平、包水平、嵌入空间三个方面对多示例学习的常见算法进行了详细的描述, 然后详细地介绍了多示例学习的算法

拓展, 为多示例学习框架移植到其他学习领域提供了一定的借鉴, 最后较为全面地介绍了多示例学习的在各个学习领域的主要应用. 基于目前的研究现状, 多示例学习已经具有十分成熟和广泛的拓展与应用, 在理论学习和社会生产实践中都发挥着十分积极的作用.

参 考 文 献

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

[2] Zhou Z H. Multi-instance learning: A survey [R]. Nanjing: Department of Computer Science & Technology, Nanjing University, 2004.

[3] Amores J. Multiple instance classification: Review, taxonomy and comparative study [J]. *Artificial Intelligence*, 2013, **201**: 81-105.

[4] Foulds J, Frank E. A review of multi-instance learning assumptions [J]. *The Knowledge Engineering Review*, 2010, **25**(1): 1-25.

[5] Chiang J Y, Cheng S R. Multiple-instance content-based image retrieval employing isometric embedded similarity measure [J]. *Pattern Recognition*, 2009, **42**(1): 158-166.

[6] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning [C]//*Proceedings of the Advances in Neural Information Processing Systems*, 2002, 561-568.

[7] Xu X. Statistical learning in multiple instance problems [D]. Hamilton: University of Waikato, 2003.

[8] Li W, Vasconcelos N. Multiple instance learning for soft bags via top instances [C]//*Computer Vision and Pattern Recognition*, 2015, 4277-4285.

[9] Xiao Y, Liu B, Hao Z, et al. A similarity-based classification framework for multiple instance learning [J]. *IEEE Transactions on Cybernetics*, 2014, **44**(4): 500-515.

[10] Li F, Sminchisescu C. Convex multiple-instance learning by estimating likelihood ratio [C]//*Proceedings of the Advances in Neural Information Processing Systems*, 2010, 1360-1368.

[11] Tax D M, Hendriks E, Valstar M F, et al. The detection of concept frames using clustering multi-instance learning [C]//*Proceedings of the International Conference on Pattern Recognition*, 2010, 2917-2920.

[12] Wang J, Zucker J D. Solving multiple-instance problem: A lazy learning approach [C]//*Proceedings of the International Conference on Machine Learning*, 2000, 1119-1126.

[13] Gärtner T, Flach P A, Kowalczyk A, et al. Multi-instance kernels [C]//*Proceedings of the International Conference on Machine Learning*, 2002, 179-186.

[14] Leistner C, Saari A, Bischof H. Miforests: multiple-instance learning with randomized trees [C]//*Proceedings of the European Conference on Computer Vision Computer Vision*, 2010, 29-42.

[15] Cheplygina V, Tax D M, Loog M. Multiple instance learning with bag dissimilarities [J]. *Pattern Recognition*, 2015, **48**(1): 264-275.

[16] Dong L. A comparison of multi-instance learning algorithms [D]. Hamilton: University of Waikato, 2006.

[17] Chen Y, Wang J Z. Image categorization by learning and reasoning with regions [J]. *The Journal of Machine Learning Research*, 2004, **5**: 913-939.

[18] Chen Y, Bi J, Wang J Z. Miles: Multiple-instance learning via embedded instance selection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(12): 1931-1947.

[19] Li W J, Yeung D Y. Mild: Multiple-instance learning via disambiguation [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(1): 76-89.

[20] Fu Z, Robles-Kelly A, Zhou J. Milis: Multiple instance learning with instance selection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(5): 958-977.

[21] Carbonneau M A, Granger E, Raymond A J, et al. Robust multiple-instance learning ensembles using random subspace instance selection [J]. *Pattern Recognition*, 2016, **58**: 83-99.
[22] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. *Nature*, 2015, **521**(7553): 436-444.
[23] Schmidhuber J. Deep learning in neural networks: An overview [J]. *Neural Networks*, 2015, **61**: 85-117.
[24] Zhu W, Hu J, Sun G, et al. A key volume mining deep framework for action recognition [C]//*Computer Vision and Pattern Recognition*, 2016, 1991-1999.
[25] Guan X, Raich R, Wong W K. Eient multi-instance learning for activity recognition from time series data using an auto-regressive hidden Markov model [C]//*Proceedings of the International Conference on Machine Learning*, 2016, 2330-2339.
[26] Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions [C]//*Proceedings of the twentieth annual symposium on Computational geometry*, 2004, 253-262.
[27] Ping W, Xu Y, Wang J, et al. Famer: Making multi-instance learning better and faster [J]. *Proceedings of the SIAM International Conference on Data Mining*, 2011, 594-605.
[28] Zhang M L, Zhou Z H. A review on multi-label learning algorithms [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(8): 1819-1837.
[29] Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning [J]. *Artificial Intelligence*, 2012, **176**(1): 2291-2320.
[30] Zhang C, Platt J C, Viola P A. Multiple instance boosting for object detection [C]//*Proceedings of the Advances in Neural Information Processing Systems*, 2005, 1417-1424.