文章编号: 1001-0920(2004)06-0607-04

多示例学习及其研究现状

蔡自兴, 李枚毅
(中南大学 信息科学与工程学院, 湖南 长沙 410083)

摘　要: 较全面地介绍和分析了第 4 种机器学习框架的多示例学习(MIL). 首先通过数学表达式对多示例学习进行描述,概括了其主要性质;然后总结了目前主要的求解多示例学习问题的算法,剖析了这些算法的主要思想;最后对多示例学习的未来发展作了展望.

关键词: 多示例学习; 测试数据集; 轴-平行矩形; 正包和负包
中图分类号: TP18　　**文献标识码**: A

Multi-instance learning and its current research status

CAI Zi-xing LI Mei-yi
(College of Information Science and Engineering, Central South University, Changsha 410083, China. Correspondent: CAI Zi-xing, E-mail: zxcai@csu.edu.cn)

Abstract: Multi-instance learning (MIL), which is regarded as the fourth learning framework, is fully introduced and analyzed. MIL is described by using mathematical formulae, and its main properties are summed up at the beginning. Then, main algorithms for solving MIL are surveyed, and the main idea of these algorithms is expounded. At the end, the future developments of MIL are presented.

Key words: multi-instance learning; test data set; axis-aligned rectangles; positive bags and negative bags

1 引　　言

多示例问题是 Dietterich 等[1]于上个世纪 90 年代中期提出的,其目的是判断药物分子是否为麝香分子(musky).

麝香分子问题是多示例学习方法的应用之一. Maron 等[2]将多示例学习方法应用于其他多示例问题,比如股票投资中的个股选择问题; Ruffo 等[3]将多示例学习方法应用于数据挖掘; Antrews 等[4], Huang 等[5], Yang 等[6], Zhang 等[7]分别将多示例学习方法用于图像检索; Chevaleyre 等[8]用多示例学习方法研究了 Mutagenesis 问题. 应用结果表明,多示例学习方法对于多示例这类不分明问题能达到较高的准确性.

多示例学习被认为是第 4 种机器学习框架,并在短短几年时间内取得了一些引人瞩目的理论成果和应用成果. 本文首先介绍多示例学习的概念,并总结出一些基本性质;然后对测试数据集 musk 进行分析,重点讨论了多示例学习的主要算法,并通过测试数据集 musk 的测试准确度对这些算法的性能进行比较;最后对多示例学习的未来发展作了展望.

2 多示例学习的概念与性质

多示例学习问题可描述为:假设训练集中每个数据是一个包(bag),每个包由一集示例(instances)组成,每个包有一个训练标记:如果包有负标记,则包中所有示例都认为是负标记;如果包有正标记,则包中至少有一个示例被认为是正标记. 学习算法需

收稿日期: 2003-05-22; **修回日期**: 2003-10-03.
基金项目: 国家自然科学基金资助项目(60234030); 教育部博士点基金资助项目(99053317).
作者简介: 蔡自兴(1939—),男,福建莆田人,教授,博士生导师,纽约科学院院士,从事智能控制、机器人学等研究;
李枚毅(1962—),男,湖南湘乡人,博士生,从事智能计算、智能机器人的研究.

要生成一个分类器,能对未知的包(unseen bags)进行正确分类.多示例学习问题可用图 1 来说明.学习算法的目标是要找出 unknown process $f(\cdot)$的最佳逼近方法.

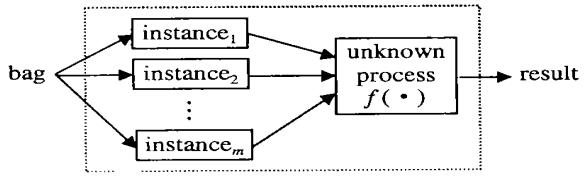

图 1 多示例学习问题描述

假设有 N 个包$\{B_1, B_2, \cdots, B_N\}$,第 i 个包由 $a(i)$ 个示例$\{B_{11}, B_{12}, \cdots, B_{1a}\}$ 组成,每个示例 B_{ij} 是一个 d 维特征向量$[B_{ij1}, B_{ij2}, \cdots, B_{ijd}]^{\mathrm{T}}$,标记集为 $\Gamma = \{l_i, l_2, \cdots, l_N \mid l_i \in \psi\}$,其中 ψ 为标记空间.记示例空间为 Ω,其子集为$\{B_1, B_2, \cdots, B_N\}$,训练数据集为

$$D = \langle B, \Gamma\rangle = \{\langle B_i, l_i\rangle \mid i = 1,2,\cdots,N\}. \tag{1}$$

定义 1 已知示例空间 Ω 及其子集(包) $B_i = \{B_{ij} \mid j = 1,2,\cdots,a(i)\}$, $i = 1,2,\cdots,N$,标记空间 $\psi = \{\text{positive}, \text{nagative}\}$,标记集 Γ 和训练数据集 $D = \langle B, \Gamma\rangle$,并且已知:

条件 1 $f_M: \{B_1, B_2, \cdots, B_N\} \rightarrow \psi$. (2)

则多示例学习问题为寻找一个映射 $\hat{f}_M$,作为真实未知映射 f_M 的最佳逼近.

如果已知包中每个示例的标记,则可计算出包的标记.于是可利用下列条件[1]:

条件 2

$$f: \{B_{ij} \mid i = 1,2,\cdots,N, j = 1,2,\cdots,a(i)\} \rightarrow \psi. \tag{3}$$

意为将 N 个包中的示例合并为一个数据集 $D_B = \{B_{ij} \mid i = 1,2,\cdots,N\}$, $j = 1,2,\cdots,a(i)$,每个数据是一个示例,可按示例学习[9,10]等方式进行学习.按这种方式对多示例问题进行学习的算法称为单示例学习算法.

条件 1 与条件 2 之间有如下关系:

命题 1 已知示例空间 Ω 及其子集(包) $B_i = \{B_{ij} \mid j = 1,2,\cdots,a(i)\}$, $i = 1,2,\cdots,N$:

1) 如果标记空间 $\psi = \{\text{TRUE(positive)}, \text{FALSE(nagative)}\}$,则对多示例问题,有

$$f_M(B_i) = f(B_{i1}) \vee f(B_{i2}) \vee \cdots \vee f(B_{ia(i)}), \quad i = 1,2,\cdots,N, \tag{4}$$

其中"$\vee$"为布尔"OR"运算;

2) 如果 ψ 是实的二值集合,正标记对应的实数比负标记大,则对多示例问题,有

$$f_M(B_i) = \max\{f(B_{i1}), f(B_{i2}), \cdots, f(B_{ia(i)})\}, \quad i = 1,2,\cdots,N. \tag{5}$$

当以示例作为训练数据时,使用条件 2 可学习到一个映射 $\hat{f}$,作为映射 f 的最佳逼近;然后按命题 1 也能构造出一个映射 $\hat{f}_M$,作为真实未知映射 f_M 的最佳逼近.

以示例或包作为训练数据,是一个利用信息量多少的问题.从这一角度可分为以下几种方式:

1) 将多示例转变为单示例,也就是只利用条件 2.这时可用各种示例学习算法(例如基于事例学习,基于实例学习,决策树算法 ID3 及其改进算法 C4.5,BP 神经网络等[9,10]) 进行学习,获取一个映射 $\hat{f}$ 作为映射 f 的最佳逼近,然后构造出映射 f_M.

2) 同时利用条件 1 和条件 2,获取一个映射 $\hat{f}$ 作为映射 f 的最佳逼近,然后构造出映射 f_M.

3) 同时利用条件 1 和条件 2,通过多示例学习算法直接获取映射 f_M.

4) 只利用条件 1,通过多示例学习算法直接获取映射 f_M.

从利用信息量看,方式 2) 和方式 3) 效果要好些,方式 1) 效果较差.文献[1]作过测试,按方式 1) 使用算法 C4.5 和反向传播 BP 神经网络的效果较差.

从研究情况划分,多示例学习算法可分成三类:一是将单示例学习算法扩展为该算法的多示例版本;二是针对多示例问题的特性构造专门的算法;三是前二者的结合,称为混合方式.

3 多示例学习的测试数据集 musk

对于多示例测试数据集的构造,通常是首先选择所讨论问题的特征向量和标记集 Γ;然后按问题要求的规则,确定一组特征向量的取值作为一个示例,若干示例组成一个包;最后按包中是否有正示例而相应地标记正包或负包.

常用的多示例测试数据集是 Dietterich 等[1]构造的 musk,其目的是通过对收集的已知分子的分析和训练学习系统,能预知一个新的分子是否可用于制药.为此,构造 24 条射线来描述分子的形状(特征向量),对每个分子结构测量相应的 24 个特征值.对于选择的两个麝香分子集(其中有 62 个重复),用计算机搜索每个分子的可能结构.然后用优化算法找出其中的低能量结构.对每个低能量结构记录其 24

个特征值, 组成表 1 所示的数据集 musk1 和 musk2(可从 UCI 机器学习数据库中提取[11]).

表 1　musk 数据集

data set	musks	non-musks	total	low energy conformation
1	47	45	92	476
2	39	63	102	6 598

从表 1 可以看出, musk1 含有 47 个正包(麝香分子) 和 45 个负包(非麝香分子), 每个包中的示例数量在 2 到 40 之间变化; musk2 含有 39 个麝香分子和 63 个非麝香分子. musk2 中的数量是 musk1 的 13.8 倍, 而分子只多了 10 个, 因此 musk2 的学习难度更大.

将表 1 中的 musk 数据集作为多示例学习算法的训练集, 每个分子作为一个包, 具有 musks 性质的分子标为正包, 没有即 non-musks 性质的分子标为负包.

4　多示例学习的主要算法

多示例学习问题的概念并不复杂, 但其求解却相当困难. 自从 1997 年 Dietterich 等[1] 发表第 1 篇多示例学习算法的文章以来, 通过近几年的研究已形成了一些有效的算法.

4.1　多示例学习成为第 4 种学习框架

多示例学习问题出现在机器学习的复杂应用中, 学习系统对每个训练例子有部分或不完全的知识, 因此多示例学习成为与监督学习、非监督学习和强化学习并列的一种机器学习, 但与监督学习和非监督学习又有所区别. 强化学习是学习"状态 →行为"的映射, 并提供了延迟奖励; 多示例学习是一类广泛存在的学习任务, 具有训练数据集的特殊性和研究方法上的特点. 包中负示例和正示例的比率(噪声比) 可能任意高, 多示例学习甚至比有噪声的监督学习更难, 这也是多示例学习需要深入研究的原因之一.

多示例学习的数据集特征介于监督学习和非监督学习之间. 从方法上说, 它可通过一定的转换机制转变为监督学习或非监督学习问题. 转变为监督学习的方法之一是将多示例问题变为单示例问题, 从而把复杂的多示例学习问题变为相对简单的单示例学习问题; 反之, 监督学习或非监督学习算法通过改造也可转变为多示例学习算法.

4.2　轴- 平行矩形分类器

Dietterich 等[1] 将每个分子视为一个包, 假定分类器可表为一个轴 - 平行矩形

$$[a_1, b_1; a_2, b_2; \cdots; a_d, b_d] = \{(x_1, x_2, \cdots, x_d) \mid a_i \leq x_i \leq b_i\}, i = 1, 2, \cdots, d.$$

并构造了几种算法来学习该矩形, 希望这个矩形能覆盖尽可能多的正示例, 且尽量不含负示例.

Long 等[8] 描述了一个多项式时间理论的算法, 并且证明如果包内的示例是从积分布中独立取的, 则轴 - 平行矩形 APR 是 PAC- 可学习的.

Auer 等[12] 证明如果包中的示例不独立, 则 APR 学习在多示例学习框架下是 NP-hard 问题, 需要使用启发式搜索方法; 并且提出一种不需要积分布的理论算法, 进而构造了相应的实际算法 MULTINST.

4.3　基于多样性密度的学习算法

Maron 等[13] 提出了多样性密度(DD) 的通用框架, 它基于如下假设: 正包减去负包的并的交点可通过最大化多样性密度来获取. 对特征空间的每个点定义一个多样性密度, 一个点附近的正包越多、负包越少, 则该点的多样性密度越大. 此算法的目标是寻找多样性密度最大的点.

Zhang 等[14] 将多样性密度与 EM 算法结合起来, 将多样性密度算法改进为 ED-DD 算法. 这是目前对 musk 数据集测试结果最好的算法.

4.4　其他机器学习的多示例版本

自多示例问题提出以来, 将其他机器学习扩充为多示例学习一直成为研究的焦点, 并取得了某些成果.

Wang 等[15] 采用 Hausdorff 距离, 扩展了 k- 最近邻算法(kNN), 用于多示例学习; Ruffo[3] 扩展了 C4.5 决断树, 构造了多示例版本 Relic.

Chevaleyre 等[8] 构造了 ID3-MI 和 RIPPER-MI 算法, 它们分别是 ID3 和 RIPPER 的多示例版本, 并进一步构造了 NAVIVE-RIPPER-MI 和 RIPPER-MI-REFINDED-COV 等算法[16].

Zhou 等[17] 通过使用特殊的误差函数反向传播神经网络, 构造了多示例学习算法 BP-MIP. 所构造的误差函数利用了多示例问题的特性, 因此算法的性能良好.

4.5　多示例神经网络和多示例 SVM

Ramon 等[18] 构造了一个多示例神经网络, 包标记和包内示例标记通过命题 1 中的 max 函数来描述. 对每个 $a(i)$ 的可能值构造了神经网络 $NN_{a(i)}$, 整个网络的输出作为 f_M 的最佳逼近. 当训练代数充分大时, 收敛点能使 f_M 与输出的平方误差充分小.

Andrews 等[4] 利用支持向量机(SVM) 来处理

多示例学习问题,称为 MI-SVM 算法,并通过人工数据和图像检索来说明该方法的有效性.

5 主要算法的性能比较

目前通用的评价算法性能的数据集只有 musk 测试数据集[1,11],因此下面仅对已提出的一些算法的测试结果进行简单的对比. 表 2 中的算法是按字母顺序排列的.

表 2 主要算法对 musk 测试数据集的性能比较

算 法	musk1 数据集 correct %	musk2 数据集 correct %
All-positive APR[1]	80.4	72.6
Backpropagation[1]	75.0	67.7
Bayesian-kNN[15]	90.2	82.4
BP-MIP[17]	83.8	未测试
C4.5[1]	68.5	58.8
Citation-kNN[15]	92.4	86.3
Diverse Density[13]	88.9	82.5
EM-DD[14]	96.8	96.0
GFS all-positive APR[1]	83.7	66.7
GFS elim-count APR[1]	90.2	75.5
GFS elim-kde APR[1]	91.3	80.4
Iterated-discrim APR	92.4	89.2
MI-SVM[4]	87.4	83.6
MULTINST[12]	76.7	84.0
Multi-instance NN[18]	88.0	82.0
Relic[3]	83.7	87.3
RIPPER-MI/ NAVIVE-RIPPER-MI[16]	88.0	77.0

从表 2 可以看出,EM-DD 算法对 musk 测试数据集的准确性最好,说明此算法抓住了多示例问题的本质特征.

6 未来发展展望

多示例学习作为第 4 种学习框架,是一类广泛存在的学习任务,它能解决许多实际问题,近几年引起了研究者的关注.

为使多示例学习具有更好的性能和更广泛的应用领域,还有不少问题需要进一步研究和讨论. 作者认为多示例学习算法的研究可从以下几个方面开展:

1) 构造新的多示例学习算法. 多示例学习的提出只有几年时间,虽然目前已形成了一些算法,但仍显得很不足,因此需要构造更多更有效的多示例学习算法.

2) 改造其他学习算法作为多示例学习算法. 多示例学习的奠基者[1]指出,一个特别有意义的问题是如何针对多示例问题修改决策树、神经网络和其他流行的机器学习算法. 相信适当地选择多示例学习的特征,构造的算法会有更好的性能.

3) 对现有多示例算法的改进. 现有的算法仍有需要改进之处,比如提高算法的准确性,针对大样本空间减少算法的计算代价等.

4) 扩展多示例学习算法的应用. 目前,该算法主要应用于药物分子活性研究和图像检索等[3,13,14,19~21]. 由于多示例问题其实是一类不分明问题,在研究和应用问题中大量存在,特别是海量数据下的检索问题已引起人们的极大兴趣,需要对多示例问题进行更深入的挖掘.

多示例学习的发展表明,多示例学习的理论研究提高了解决多示例问题的能力;反过来,多示例学习的应用也促进了多示例学习理论的深入研究. 可以预见,在最近几年内,将有更新的多示例学习算法和更多的多示例学习应用成果.

参考文献(References):

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

[2] Maron O, Ratan A L. Multiple-instance learning for natural scene classification[A]. *Proc of the 15th Int Conf on Machine Learning*[C]. Madison, 1998. 341-349.

[3] Ruffo G. Learning single and multiple instance decision tree for computer security applications[D]. Torino: University of Turin, 2000.

[4] Andrews S, Hofmann T, Tsochantaridis I. Multiple instance learning with generalized support vector machines[A]. *AAAI/ IAAI*[C]. Edmonton, 2002. 943-944.

[5] Huang X, Chen S C, Shy M L, et al. User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval[A]. *MDM/ KDD2002 Workshop*[C]. Edmonton, 2002. 100-108.

[6] Yang C, Lozano P T. Image database retrieval with multiple-instance learning techniques[A]. *Proc of the 16th Int Conf on Data Engineering*[C]. San Diego, 2000. 233-243.

[7] Zhang Q, Goldman S A, Yu W, et al. Content-based image retrieval using multiple-instance learning[A]. *The Nineteenth Int Conf on Machine Learning*[C]. Sudney, 2002. 682-689.

(下转第 615 页)

Tampa ,1998. 3757-3762.

[2] Mignone D, Bemporad A, Morari M. A framework for control, fault detection, state estimation and verification of hybrid systems[A]. *Proc of American Control Conf*[C]. San Diego ,1999. 134-138.

[3] Koutsoukos X, Zhao F, Haussecker H, et al. Fault modeling for monitoring and diagnosis of sensor-rich hybrid systems[A]. *Proc of the 40th IEEE Conf on Decision and Control*[C]. Orlando ,2001. 793-801.

[4] Michael W H, Brian C W. Mode estimation of probabilistic hybrid systems[A]. *Hybrid Systems: Computation and Control*[C]. Berlin: Springer-Verlag ,2002. 253-266.

[5] Thomas Henzinger. The theory of hybrid automata[A]. *Proc of the 11th Annual IEEE Symposium on Logic in Computer Science*[C]. New Brunswick ,1996. 278-292.

[6] Gordon N J, Salmond D J, Smith A F M. Novel approach to nonlinear/non — Gaussian-Bayesian state estimation [J]. *IEE Proceedings F* ,1993 ,140(2) :107-113.

[7] Doucet A, Gordon N, Krishnamurthy V. Particle filters for state estimation of jump Markov linear systems[J]. *IEEE Trans on Signal Processing* ,2001 ,49(3) :613-624.

[8] Casella G, Robert C P. Rao-blackwellisation of sampling schemes[J]. *Biometrika* ,1996 ,83(1) :81-94.

[9] Liu J S, Chen R. Sequential Monte-Carlo methods for dynamic systems [J]. *J of the American Statistical Association* ,1998 ,93 :1032-1044.

[10] Mo Y W, Xiao D Y. Hybrid system monitoring and diagnosing based on particle filter algorithm [J]. *Acta Automatica Sinica* ,2003 ,29(4) :641-648.

(上接第 610 页)

[8] Chevaleyre Y, Zucker J D. Solving multiple-instance and multiple-part learning problems with decision trees and decision rules: Application to the mutagenesis problem[A]. *Lecture Notes in Artificial Intelligence* [C]. Berlin: Springer ,2001. 204-214.

[9] Tom M. **曾华军, 张银奎译. 机器学习**[M]. **北京:机械工业出版社** ,2003.

[10] **蔡自兴,徐光祐. 人工智能及其应用**[M]. **第 2 版. 北京:清华大学出版社** ,1996.

[11] Blake C L, Keogh E, Merz C J. UCI repository of machine learning databases[R]. Irvine: University of California ,1998.

[12] Auer P. On learning from multi-instance examples: Empirical evaluation of a theoretical approach[A]. *Proc of the 14th Int Conf on Machine Learning*[C]. Nashville, 1997. 21-29.

[13] Maron O, Lozano P T. A framework for multiple-instance learning[A]. *Advances in Neural Information Processing Systems*[C]. Cambridge: MIT Press ,1998. 570-576.

[14] Zhang Q, Goldman S A. EM-DD: An improved multiple-instance learning technique[A]. *Advances in Neural Information Processing Systems* [C]. Cambridge: MIT Press ,2002. 1073-1080.

[15] Wang J, Zucker J D. Solving the multiple-instance problem: A lazy learning approach[A]. *Proc of the 17th ICML*[C]. San Francisco ,2000. 149-166.

[16] Chevaleyre Y, Bredeche N, Zucker J D. Learning rules from multiple instance data: Issues and algorithms[A]. *The 9th Int Conf on Information Processing and Management of Uncertainty in Knowledge-based Systems*[C]. Annecy ,2002. 117-124.

[17] Zhou Z H, Zhang M L. Neural networks for multi-instance learning[A]. *Proc of the Int Conf on Intelligent Information Technology*[C]. Beijing ,2002. 455-459.

[18] Ramon J, Raedt L D. Multi-instance neural networks [A]. *Proc of ICML-2000 Workshop on Attribute-value and Relational Learning*[C]. Stanford ,2000. 53-60.

[19] Scott S. Geometric patterns: Algorithms and applications [A]. *ICML 2000 Workshop on Machine Learning of Spatial Knowledge*[C]. *Palo Alto* ,2000. 109-115.

[20] Auer P, Long P M, Srinivasan A. Approximating hyper-rectangles: Learning and pseudo-random sets[J]. *J of Computer and System Sciences* ,1998 ,57(3) :376-388.

[21] Blum A, Kalai A. A note on learning from multiple instance examples[J]. *Machine Learning* ,1998 ,30(1) :23-29.