ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 2,Mar. 2018,pp. 359－369
DOI:10.16337/j.1004-9037.2018.02.019

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

多示例学习的示例层次覆盖算法

董露露[1]　谢　飞[2]　章　程[3]

(1.安徽广播电视大学安徽继续教育网络园区管理中心，合肥，230022;2.合肥师范学院计算机科学与技术系,合肥,230601;3.安徽大学计算机科学与技术学院,合肥,230039)

摘　要：在多示例学习(Multi-instance learning，MIL)中,核心示例对于包类别的预测具有重要作用。若两个示例周围分布不同数量的同类示例,则这两个示例的代表程度不同。为了从包中选出最具有代表性的示例组成核心示例集,提高分类精度,本文提出多示例学习的示例层次覆盖算法(Multi-instance learning with instance_level covering algorithm，MILICA)。该算法首先利用最大 Hausdorff 距离和覆盖算法构建初始核心示例集,然后通过覆盖算法和反验证获得最终的核心示例集和各覆盖包含的示例数,最后使用相似函数将包转为单示例。在两类数据集和多类图像数据集上的实验证明,MILICA 算法具有较好的分类性能。

关键词：多示例学习;覆盖算法;核心示例集;相似度函数

中图分类号：TP18　　**文献标志码**:A

Multi-instance Learning with Instance-Level Covering Algorithm

Dong Lulu[1], Xie Fei[2], Zhang Cheng[3]

(1. Management Centre of Anhui Continuing Education Online Campus, Anhui Radio and TV University, Hefei, 230022,China; 2. Department of Computer Science and Technology, Hefei Normal University, Hefei,230601,China;3. College of Computer Science and Technology, Anhui University,Hefei, 230039,China)

Abstract: In multi-instance learning, the core instances play an important role on the prediction of bags′ label. And if two instances have different numbers of instances with the same category around them, they have different levels of representative. In order to improve the classification accuracy, multi-instance learning with instance-level covering algorithm (MILICA) is proposed by which we could select the most representative instances to form the core instance set. Firstly, with the max Hausdorff distance and the covering algorithm, the initial core instance set is constructed. Then, the final core instance set and the number of instances in a cover are obtained. Finally, a similarity measure function is used to convert a bag into a single sample for classification. Experimental results on two-category datasets and multi-category image datasets demonstrate that the proposed MILICA method has perfect classification capability.

Key words: multi-instance learning; covering algorithm; core instance set; similarity measure function

基金项目:国家自然科学基金(61503116)资助项目;安徽省教育厅自然科学基金重点(KJ2014A081)资助项目;安徽省级优秀青年基金重点(2013SQRL097ZD)资助项目;安徽省自然科学基金(1408085QF108)资助项目。
收稿日期:2017-03-30;**修订日期**:2017-10-30

引 言

多示例学习这一新型机器学习框架是 Dietterich 等于 1997 年进行药物分子活性预测研究时提出的[1]。其实质是对由多个示例组成的包进行学习并对未知标记的包进行预测。目前已在图像分类[2]、图像检索[3-5]、视觉追踪[6]和行人检测[7]等方面得到广泛的应用。

总体来说,多示例学习主要分为两类。一类从包与示例之间的关系出发,寻求解决多示例学习问题的途径。1988 年,Maron 等[8]提出多样性密度(Diverse density, DD)算法。DD 算法通过多次梯度下降搜索来求解多样性密度点,但该算法计算时间较长,效率不高,且并不能确保找到全局最优解。Zhang 等[9]在 DD 算法的基础上,引入期望最大化(Expectation maximization, EM)算法,提出期望最大多样性密度算法 (Expectation maximization vision of diverse density, EM-DD)。但该算法要通过不断迭代获取多样性密度最大的示例,且正包中的正示例可能是随机分散的,选出的目标示例不一定能有效代表所有的正示例,因而会影响分类效果。

另一类则通过对传统的单示例学习方法进行改进来解决多示例问题[10]。该类方法主要分为两种,一种是直接为示例加上对应的包的标记来解决多示例问题[11]。例如 Andrews 等[12]将支持向量机(Support vector machine,SVM)引入多示例学习,并提出 MI-SVM 和 mi-SVM 算法;Zhou 以半监督学习的视角看待多示例学习问题,并提出基于特殊半监督支持向量机方法的多示例学习(Multiple instance learn with semi-super vised SVM,MissSVM)[13]算法;Shao 等[14]对双支持向量机进行了扩展,提出多示例双支持向量机(Multi-instance twin support vector machines, MI-TWSVM)算法;Qi 等[15]通过选出正包中很大可能属于正示例的示例构造非平行分类器,提出 MI-NSVM 算法。由于正包中可能存在大量伪正例,这种方法很难有效解决多示例学习问题。另一种方法则通过提取核心示例集,将包转化为用特征向量表示的单示例,进而使用传统监督算法学习。例如 Chen 等[16]使用 DD 算法提取每个包中具有最大多样性密度的示例,提出 DD-SVM 算法;Chen 等[17]基于一种新型特征映射方法和 1-norm 支持向量机模型,提出基于嵌入式示例选择的多示例学习(Multiple instance learning via embedded instance selection, MILES)算法,在实现特征提取的同时完成分类;Li 等[18]利用基于示例类别消歧的方法提取核心示例,提出 MILD_ B算法;Fu 等[19]提出基于示例选择的多示例学习(Multiple instance learning with instance selection,MILIS)算法,该算法使用核密度估计方法选取核心示例,并通过一种最优化框架来构建分类器;Erdem 等[20]通过对包中的示例建立主导集找到核心示例,提出基于主导集的多示例学习(Multiple-instance learning with instance selection via dominant sets,MILDS)算法;Li 等[21]基于多核框架将每个图像包转为单示例,并使用多核支持向量机(Multiple-kernels support vector machine,MKSVM)进行分类,提出 MKSVM-MIL 算法。上述关于提取核心示例集的方法并未考虑所提取示例的代表程度,因而影响了分类效果。为此,本文提出一种新的基于示例提取的多示例层次覆盖算法(Multi-instance learning with a instance-level coering algorithm, MILICA)。该算法利用覆盖算法[22](Covering algorithm, CA)选出正负包中具有代表性的示例,并使用覆盖的示例数表示所提取示例的代表程度,每个覆盖可视为一个聚类,覆盖中心即为聚类的中心。

以上多示例学习的研究主要集中于多示例分类问题。近年来,非监督多示例聚类问题也开始受到研究人员的关注。Zhang 等[23]率先对非监督多示例学习问题进行研究,提出一种包级多示例聚类算法(Bag-level multi-instance clustering,BAMIC),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。基于 BAMIC 的聚类结果,又提出一种基于包级转换表示的多示例预测算法(Bag-level representation transformation for multi-instance prediction, BARTMIP)。Zhang[24]提出一种用于多示例聚类的新的框架 M^3IC(Maximum margin multiple instance clustering, M^3IC)。M^3IC 致力于在每个包的至少一

个示例上找到最大化边缘差距来实现多示例聚类。上述两篇文献的研究重点在于多示例聚类,尤其文献[24]专门针对多示例聚类问题构建了一种新的框架。而本文的研究目标是得到一种用于多示例分类的算法,并尽量确保算法能得到较高的分类准确度,重点是如何抽取出最具有代表性的核心示例。实验结果表明,相比于已有的多示例学习算法,该算法有效提高了分类准确率。

1 相关定义和算法

1.1 多示例学习

在多示例学习中,训练集样本是一个个包,包具有类别标记,每个包由若干没有类别标记的示例组成。当一个包至少包含一个正示例时,称该包为正包,将其包含的非正例称为假正例。当一个包中的示例均是负示例,则称该包为负包。多示例学习的难点在于训练集中包的标记已知,而示例标记未知,获取的信息有限,且正包中含有大量的假正例,从而增加了学习的复杂性[25-26]。多示例学习的目标是通过对训练包进行学习构造基于示例的分类器 $f(x):x\to y,x\in\chi$ 或者基于包的分类器 $F(X):X\to y$ 对未标记的包进行分类[1-6,27-33]。具体框架如图 1 所示。

1.2 覆盖算法

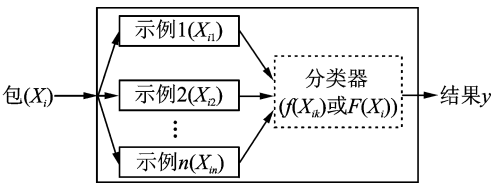

图 1 多示例学习框架

Fig. 1 Framework of MIL

张铃、张钹教授提出的基于覆盖的构造性机器学习方法,简称覆盖算法[22]。在此给出该算法的完整过程:给定样本集 $D=\{(\boldsymbol{v}_i,l_i)\mid i=1,\cdots,m;l_i=1,\cdots,j\}$,其中 $\boldsymbol{v}_i$ 和 l_i 分别表示第 i 个样本和该样本的类别,$\boldsymbol{v}_i$ 是 d 维特征向量,m 表示样本的数量,j 表示样本的类别数。定义 $V=\{V_1,V_2,\cdots,V_t\}$ 为根据类别划分的样本集合,其中 $V_i\in V(i=1,\cdots,j)$ 为第 i 类的样本子集。定义 $C=\{c_i\mid c_i=(\text{center}_i,r_i,\text{no}_i),i=1,2,\cdots\}$ 表示通过覆盖算法得到的球形领域覆盖集,其中 center_i 表示覆盖 c_i 的覆盖中心,覆盖中心可作为覆盖的代表性样本,r_i 和 no_i 分别表示 c_i 的覆盖半径和覆盖样本数。用 $\text{flag}(\boldsymbol{v})$ 表示样本 $\boldsymbol{v}$ 是否属于某覆盖,若 $\text{flag}(\boldsymbol{v})=1$,则表示 $\boldsymbol{v}$ 落入了某一覆盖 c_i 中。用$<\boldsymbol{v},\boldsymbol{v}'>$表示样本 $\boldsymbol{v}$ 和 $\boldsymbol{v}'$ 的内积,该值与两样本间的欧氏距离呈反比。算法 1 给出了 CA 的训练过程。

算法 1 CA 的训练过程

输入:训练集 D_1

输出:覆盖集 C

(1)初始化:令 $C=\varnothing$,$\text{flag}(v_i)=0,i=1,\cdots,m$,同时定义转换函数 $T(\boldsymbol{v})=(\boldsymbol{v},\sqrt{R^2-\|\boldsymbol{v}\|^2})$,$R=\max\{\|\boldsymbol{v}\|\boldsymbol{v}\in V\}$;

(2)使用步骤(1)中的转换函数构造映射:$\boldsymbol{v}\to\boldsymbol{H}^d$,其中 $\boldsymbol{H}^d$ 是 $d+1$ 维样本空间的一个 d 维球面,据此产生一个新的训练集 D_1;

(3)对每一个 $V_i\subseteq D_1$,随机选取一个样本 $\boldsymbol{v}\in V_i$ 且 $\text{flag}(\boldsymbol{v})=\text{false}$,计算

$$d_1=\max\{<\boldsymbol{v},\boldsymbol{v}'>\mid\boldsymbol{v}\in V_i,\boldsymbol{v}'\notin V_i\}$$

$$d_2=\min\{<\boldsymbol{v},\boldsymbol{v}'>\mid\boldsymbol{v},\boldsymbol{v}'\in V_i,<\boldsymbol{v},\boldsymbol{v}'>>d_1\}$$

$$r=(d_1+d_2)/2$$

以 $\boldsymbol{v}$(即 $\text{center}=\boldsymbol{v}$)为覆盖中心,以 r 为半径,构造覆盖 c。之后,计算落入该覆盖的样本数 no。令 $\text{flag}(\boldsymbol{v})=1$ 和 $\text{flag}(\boldsymbol{v}'')=1$,此处 $\boldsymbol{v}''$属于覆盖 c。将覆盖 c 加入 C,重复执行上述操作直到对任意 $\boldsymbol{v}\in V_i$,$\text{flag}(\boldsymbol{v})=1$。

可见,CA 通过构造一个 3 层的前向神经网络获得覆盖集 C。该神经网络将样本向量作为输入,将

C 中的各个覆盖作为隐层节点，其输出层则将隐层输出采取或门的方式连接，至此即可得到输入样本的标记。由于每个覆盖包含的样本均同属于一个类别，可将一个覆盖视为一个聚类，从而可以使用覆盖的半径来衡量相应聚类的范围大小，可将覆盖中心视为聚类中心，并根据覆盖包含的样本数量来衡量聚类中心的代表程度。

对覆盖 $c_i \in C$ 和测试样本 $\boldsymbol{t}_s$，神经网络的隐层输出[15,16-31]定义为

$$\text{result}_i = \text{sign}(< \boldsymbol{t}_s, \text{center}_i > - r_i) \cdot l_i \tag{1}$$

式中：sign 是符号函数，l_i 表示 c_i 的覆盖中心类别。若 $\text{result}_i > 0$，则表示 $\boldsymbol{t}_s$ 落入 c_i 中，即被 c_i 覆盖。若某测试样本未被任何覆盖包含或同时被不少于两个的覆盖包含，则采用类似于聚类的方式，将其并入距离最近的覆盖。

2 MILICA 算法

该算法先通过距离函数 Hausdorff 和 CA 获取具有一定代表性的核心示例，再利用相似度函数将每个包转为单示例，最后使用 CA 进行学习和测试。

2.1 符号说明

定义 χ 表示示例空间，数据集 $D_S = \{(X_1, y_1), (X_2, y_2), \cdots, (X_m, y_m)\}$，其中第 ii 个包定义为 $X_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{in_i}\} \subseteq \chi, i = 1, \cdots, m, y_i \in \{-1, +1\}$ 为对应的类别，n_i 表示该包中的示例数，$\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \cdots, x_{ijd}]$，即 x_{ij} 由一个 d 维的特征向量组成。定义 $X = \{X_1^+, \cdots, X_{m^+}^+, X_1^-, \cdots, X_{m^-}^-\}$ 表示含有 m^+ 个正包和 m^- 个负包的样本集，其中 X_i^+, X_i^- 分别表示第 ii 个正包和第 ii 个负包。令 $X^+ = \{X_1^+, \cdots, X_{m^+}^+\}$ 与 $X^- = \{X_1^-, \cdots, X_{m^-}^-\}$ 分别代表正包和负包的集合。覆盖集用 $C_S = \{c_i \mid c_i = (x_i, r_i, no_i) \mid x_i \in X, i = 1, 2, \cdots; j = 1, 2, \cdots, m\}$ 表示，其中 x_i 为覆盖 c_i 的覆盖中心，r_i 为 c_i 的覆盖半径，no_i 为被 c_i 覆盖的示例数。CIS^+ 和 CIS^- 分别表示提取的正包和负包中的核心示例集。

2.2 核心示例集提取

2.2.1 两类问题的核心示例提取

首先，从每个正包中选出一个示例组成初始核心示例集。由于负包中的示例均为负示例，可使用距离函数度量正包与负包之间的差异性，进而选出若干正示例。本文使用最大 Hausdorff 距离[5]，定义正包 X_j^+ 到负包集 X^- 的最大 Hausdorff 距离为

$$d(X_j^+, X^-) = \max_{\boldsymbol{x}_k \in X^-} \min_{\boldsymbol{x}_j \in X_j^+} \| \boldsymbol{x}_j - \boldsymbol{x}_k \|^2 \tag{2}$$

式中 $\| \boldsymbol{x}_j - \boldsymbol{x}_k \|^2$ 表示示例 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_k$ 的欧式距离。从 X_j^+ 选取具有代表性的核心示例为

$$x_j^+ = \arg d(X_j^+, X^-) \tag{3}$$

按上述方法选出的各正示例组成了最初的核心示例集 CIS^+，$\text{CIS}^+ = \{x_1^+, x_2^+, \cdots, x_{m^+}^+\}$。

然后，通过 CA 和反验证逐步排除正包中的假正例。先将 X^- 和 CIS^+ 中的示例分别标记为 -1 和 $+1$，并通过 CA 对 X^- 和 CIS^+ 求覆盖，将获得的负示例和正示例的覆盖中心集分别记为和。再使用覆盖中心为 XC^- 的覆盖对正包中的示例进行反验证，选出其中未被覆盖且不属于 CIS^+ 的示例组成示例集 NC。对于 NC 中的每一个示例 nx_i，根据式(4)计算其与 CIS^+ 中所有示例之间的距离 d，并选出最小距离作为 d。

$$d(nx_i, x_j) = \min_{x_j \in \text{CIS}^+} \| n\boldsymbol{x}_i - \boldsymbol{x}_j \|^2 \tag{4}$$

对 d 由小到大进行排序，并据此调整对应示例在 NC 中的位置。即对 NC 中任意两个示例 nx_i，nx_k，若 $d(nx_i, x_j) < d(nx_k, x_j)$，则将 nx_i 排在 nx_k 之前，调整后 $\text{NC} = \{nx_1, nx_2, \cdots, nx_h\}$，其中 h 为 NC 中的示例数。

从 NC 中选出前 $g*h$ 个示例并入 CIS^+,而剩余距离较远的示例可能是假正例,将其并入 X^-,从而逐步排除正包中的假正例。当 $g=0$ 时,将 NC 中的示例全部归为 X^-,当 $g=1.0$ 时,将 NC 中的示例全部归为 CIS^+。

最后,从正包和负包中选出具有代表性的示例组成最终核心示例集 CIS。通过 CA 对 CIS^+ 与 X^- 重新求覆盖得到正示例的覆盖集 $CS^+=\{c_i^+ \mid c_i^+=(x_i^+,r_i^+,no_i^+),i=1,2,\cdots,k^+\}$ 和负示例的覆盖集 $CS^-=\{c_i^- \mid c_i^-=(x_i^-,r_i^-,no_i^-),i=1,2,\cdots,k^-\}$,其中 k^+ 和 k^- 分别为 CS^+ 和 CS^- 中的覆盖中心数。由 CS^+ 和 CS^- 中的覆盖中心分别组成集合 CIS^+ 和 CIS^-,最终获得核心示例集 $CIS=\{CIS^+,CIS^-\}$,示例数集 $no=\{no_1^+,\cdots,no_{k^+}^+,no_1^-,\cdots,no_{k^-}^-\}$。

图 2 给出两类样本的 CIS 提取过程。其中,小圆圈和等边三角形分别表示正包和负包中的示例。通过 5 个满足协方差矩阵为单位矩阵的正态分布 $N_1\sim N([5,5]^T,\boldsymbol{I})$,$N_2\sim N([5,-5]^T,\boldsymbol{I})$,$N_3\sim N([-5,5]^T,\boldsymbol{I})$,$N_4\sim N([-5,-5]^T,\boldsymbol{I})$,$N_5\sim N([0,0]^T,\boldsymbol{I})$得到各个示例,这里 $N([5,5]^T,\boldsymbol{I})$表示该正态分布的平均值为$[5,5]^T$。一个包为正包当且仅当其包含的示例源自于 N_1,N_2,N_3 当中的至少两个正态分布,否则该包为负包。图 2~3 均包含 6 个正包和 6 个负包,分别使用数字 1~6 和 7~12 来标记,每个包最多包含 8 个示例,且示例与其所在包具有同样的标记。

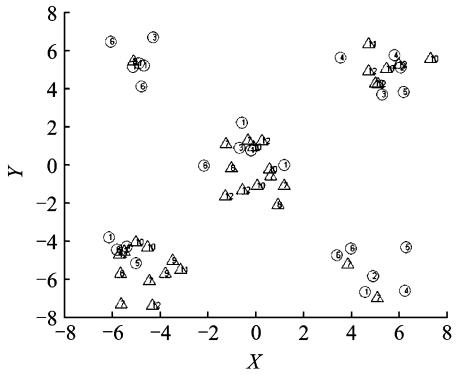

图 2 原始示例分布

Fig. 2 Raw instance distributions

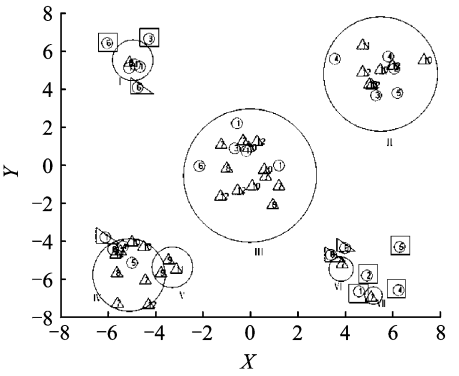

图 3 两类问题对应的 CIS 提取过程

Fig. 3 CIS extraction process for two-category problem

图 3 中长方形表示根据式(3)从正包中选取的初始示例,三角形表示最终获得的 CIS^+,大的空心圆表示负示例覆盖集 CS^-,用Ⅰ~Ⅶ标记。

从图 2 可看到负包中的示例周围分布着很多正包中的示例,它们很可能是假正例,应尽可能将其从正包中排除。从图 3 可看到选取的初始示例周围有较少负示例,则这些初始示例很大可能是正示例,应首先将其选出。此外,覆盖Ⅲ包含的示例最多,这就使得对于任意一个随机选取的测试示例而言,其被覆盖Ⅲ包含的概率最大,因此其覆盖中心的代表程度也就最大。

2.2.2 多类问题的核心示例提取

2.2.1 节中的方法可直接用来处理两类问题的多示例学习,但它采用的 one-vs-rest 策略并不适用于多类问题的多示例学习。one-vs-rest 策略选取某一类作为正包类,剩余的几类作为负包类且其中的示例全部认为是负示例。但在多示例学习中,不同类别的负包中的示例之间并无直接关系,不能简单将其作为一类利用 CA 算法进行聚类求覆盖[27]。所以在多类问题中本节采用如下方法。

对于给定的 n 类包,首先将第 i 类作为正包类,其余几类作为负包类,通过式(3)从每个正包中选出一个示例组成第 i 类的初始示例集 CIS^i。然后利用 CA 对 CIS^i 与负包中的负示例求覆盖,求覆盖时应将负包中的示例按其原始类别单独求覆盖,而非将其作为一类求覆盖。接着使用式(4)和反验证从第 i

类正包中选出未被负类覆盖且不属于 CIS^i 的示例，这些示例组成的集合用 NC_i 表示。采用 2.2.1 节中的方法调整 NC_i 中示例的位置，并根据参数 g 更新 CIS^i。最后使用 CA 对 CIS^i 与负包中的示例求覆盖，将得到的第 i 类的覆盖中心作为最终的 CIS^i。重复上述步骤，依次提取其余几类的 CIS^i 得到 CIS

$$CIS=\bigcup_{i=1}^{n}CIS^i=\{x_1^1,\cdots,x_{m^1}^1,\cdots,x_1^n,\cdots,x_{m^n}^n\} \tag{5}$$

式中：CIS^i 表示从第 i 类包提取的核心示例集，m^i 表示从第 i 类包提取得到的核心示例的个数，且这些核心示例对应的覆盖的示例数集为 $no=\{no_1^1,\cdots,no_{m^1}^1,\cdots,no_1^n,\cdots,no_{m^n}^n\}$。

2.3 转换和分类

由 2.2 节得到的核心示例集 CIS 和覆盖的示例数集 no 定义一个相似度函数，即计算 x_k 与包 X_i 中距离最近的示例之间的相似度，则

$$s(x_k,X_i)=\begin{cases}(no_k/\sum_{p=1}^{k^+}no_p)*\min\limits_{x_{ij}\in X_i,x_k\in CIS^+}\exp(-\dfrac{d(x_k,x_{ij})^2}{2\sigma^2})\\(no_k/\sum_{p=1}^{k^-}no_p)*\min\limits_{x_{ij}\in X_i,x_k\in CIS^-}\exp(-\dfrac{d(x_k,x_{ij})^2}{2\sigma^2})\end{cases} \tag{6}$$

式中参数 σ 与核心示例间的平均距离有关，exp 为以自然数为底的指数函数。

对于给定的包 X_i，在得到其与 x_k 的相似度之后，根据式(7)将其表示为一个单示例。该单示例是一个 k^++k^- 维的特征向量，且其标记和包的标记相同。

$$\boldsymbol{\varphi}(X_i)=[s(x_1^+,X_i),\cdots,s(x_{k^+}^+,X_i),s(x_1^-,X_i),\cdots,s(x_{k^-}^-,X_i)]^{\mathrm{T}} \tag{7}$$

若 X_i 为正包，则 x_i^+ 与 X_i 之间的相似度较大，x_i^- 与 X_i 的相似度较小，即 X_i 在 $\boldsymbol{\varphi}(X_i)$ 中的各分量所占的权重不同。

对于多类问题，由式(4,5)得到 CIS 和 no 后，定义转换函数为

$$\boldsymbol{\varphi}(X_i)=[s(x_1^1,X_i),\cdots,s(x_{m^1}^1,X_i),s(x_1^n,X_i),\cdots,s(x_{m^n}^n,X_i)]^{\mathrm{T}} \tag{8}$$

与两类问题一样，包的标记与其转换后得到的示例的标记相同。完成上述转换后，就可以用 CA 进行多示例分类了。

算法 2 给出两类问题的 MILICA 算法的训练过程。因为多类问题仅在 CIS 的提取上有所不同，其余过程和两类问题一致，所以文中并未介绍其训练过程。

算法 2 MILICA 算法的训练过程

输入：训练包 $X=\{X_1^+,\cdots,X_{m^+}^+,X_1^+,\cdots,X_{m^-}^+\}$

输出：CIS 和 CS

(1)将训练集中的正包(负包)标记为+1(−1)；

(2)利用公式(3)从正包中选出 m^+ 个正示例组成初始的 CIS^+；

(3)通过 CA 对 CIS^+ 和负包中的全部示例求覆盖，得到覆盖中心集合和，并利用对正包中的示例进行反验证，更新和；

(4)对和 X^- 重新求覆盖，得到 CIS 和 no，然后利用式(6)和式(7)将每个包转为单示例，其标记与包一致；

(5)对转换后的示例求覆盖得到覆盖集 CS。

测试时，先利用式(6,7)将每个测试包转换为单示例，之后利用学习过程得到的 CS 进行测试。

2.4 时间和空间复杂度

从上述算法步骤可以看出，MILCA 需要根据式(2,3)计算每个示例与所有示例之间的距离以构建核心示例，根据式(6,7,8)计算核心示例与各个包中距离最近示例之间的相似度以便对包进行转换。因

此,MILCA 的时间和空间复杂度均为 $O(e^2)$,其中 e 为总的示例数。可见,本文算法与 DD,EM-DD 和 DD-SVM 等经典算法相比,未产生额外的时间和空间开销。

3 实验结果和分析

本文通过 2 组不同的数据集对 MILICA 算法进行了实验。第 1 组使用仅包含正包和负包的两类数据集,第 2 组使用多类图像数据集。实验中的参数 $\sigma=\text{linspace}(0.6\mu,1.4\mu,10)$。其中,$\text{linspace}(a,b,n)$ 表示在 a 和 b 之间等间隔的取 n 个值,包括 a 和 b。μ 为 CIS 中每两个示例之间欧式距离的平均值,本实验中 μ 的取值分别为 5,10,15,20,25,30,35,40,45 和 50。$g=0,0.2,0.4,0.6,0.8$ 和 1.0,即每次选取 6 个不同的 g 值进行实验。

3.1 两类数据集分类

本节使用多示例学习中 5 个常用数据集衡量 MILICA 算法,分别为 Musk1,Musk2,Elephant,Fox,Tiger,其中 Musk1 和 Musk2 是药物分子活性预测数据集[1],Elephant,Fox 及 Tiger 分别代表大象、狐狸和老虎 3 类不同的多示例图像集。表 1 给出 5 个数据集的详细信息。

实验采用 10 次 10-fold CV(10 交叉验证)方法进行,且每次交叉验证均通过调整 σ 得到不同的核心示例集,最终结果即 10 次交叉验证结果的算数平均值。表 2 给出不同 g 值下 MILICA 算法得到的平均分类准确度。为验证本文所提算法 MILICA 的有效性,将其与 MILDS,MILIS,DD-SVM 等经典算法的实验结果进行对比分析。表 3 给出了 MILICA 算法在 5 个数据集上得到的平均份额里准确度及 95% 的置信区间,同时也给出了其他 12 个算法的分类结果。对比算法中,除 MILIS 使用 15 次 10-fold CV(10 交叉验证)外,其余均采用 10 次 10-fold CV。对比算法的结果按表 3 中从上到下的顺序分别取自文献[8,9,12~14,16,17,19,20,34]。表 3 按照算法是否基于核心示例提取的思想分为 3 部分,其中,核心示例提取算法的分类结果位于第 2~7 行,非示例提取算法的分类结果位于第 8~12 行,除本文算法 MILICA 之外的其余算法在各个数据集上平均分类准确度的均值位于最后一行。表 2 中当 $g=0.2$ 时,MILICA 在 Tiger 数据集上取得 85.1% 的准确度;$g=0.6$ 时,MILICA 在 Musk2 数据集上取得91.4% 的准确度;$g=0.8$ 时,MILICA 在 Fox 数据集上取得 65.9% 的准确度;$g=1.0$ 时,MILICA 在 Elephant 数据集上取得 85.0% 的准确度。

从表 3 可以看出,MILICA 算法在 5 个数据集上的准确度比其他 MIL 算法的平均值都至少高出 2.4%,特别是在 Tiger 上更是高出 6.6%,相应的 95% 的置信区间也小于其他算法。其中,在 Musk1 上,MILICA 算法的分类准确度为 90.7%,仅低于 MI-TSVM,明显高于其他示例提取算法。在其余 4 个数据集上,MILICA 都取得最高的分类准确率。非示例提取算法 DD,mi-SVM,MI-SVM,EM-DD 和 MissSVM 在 5 个数据集上的准确度均低于 MILICA。

表 1 两类数据集的详细信息

Tab. 1 Detail information of two-category datasets

数据集	正包/负包数	平均示例数/包	维数
Musk1	47/45	5.17	166
Musk2	39/63	64.69	166
Elephant	100/100	6.96	230
Fox	100/100	6.60	230
Tiger	100/100	6.10	230

表 2 不同 g 值下 MILICA 算法在标准数据集上的平均分类准确度

Tab. 2 Average classification accuracy of MILICA over different values of g on benchmark datasets %

g	Musk1	Musk2	Elephant	Fox	Tiger
0	90.7	88.6	83.9	61.2	80.5
0.2	87.9	86.5	82.3	58.5	85.1
0.4	88.6	88.0	78.7	58.1	77.3
0.6	88.4	91.4	80.2	60.4	78.0
0.8	87.6	87.8	83.6	65.9	77.1
1.0	88.6	87.6	85.0	62.6	75.9

表 3 不同 MIL 算法在标准数据集上的分类结果比较

Tab. 3 Comparison of the performance of various MIL algorithms on benchmark datasets %

算法	Musk1	Musk2	Elephant	Fox	Tiger
MILICA	90.7:[89.8,91.1]	91.4:[90.4,92.1]	85.0:[84.6,85.5]	65.9:[64.8,66.2]	85.1:[84.4,86.2]
MI-TWSVM	93.6	88.2	83.5	62.5	79.0
MILDS	75.0	86.1	84.8	64.3	81.5
MILD_B	88.3:[86.2,89.6]	86.8:[85.1,87.9]	82.9:[81.6,83.9]	55.0:[54.4,56.3]	75.8:[74.3,76.5]
MILIS	88.6:[85.8,91.5]	91.1:[89.4,92.8]	*n/a*	*n/a*	*n/a*
MILES	86.3:[84.9,87.0]	87.7:[86.3,89.1]	84.1:[82.8,85.3]	63.0:[62.4,65.2]	80.7:[78.6,82.9]
DD-SVM	85.8:[84.1,87.6]	91.3:[90.4,92.8]	83.5:[82.3,85.2]	56.6:[55.2,58.0]	77.2:[76.5,78.7]
mi-SVM	87.4:[85.7,89.3]	83.6:[82.2,84.8]	82.0:[80.6,83.5]	58.2:[56.8,59.6]	78.9:[80.2,80.4]
MI-SVM	77.9:[76.7,79.0]	84.3:[82.8,85.8]	81.4:[78.3,84.5]	59.4:[58.2,60.5]	84.0:[82.5,85.6]
DD	88.9	82.5	83.0	65.5	74.0
EM-DD	84.8	84.9	78.3	56.1	72.1
MissSVM	87.6	80.0	*n/a*	*n/a*	*n/a*
MI-ELM	88.2	85.0	82.6	62.7	84.6
Average	85.7	78.0	82.6	60.3	78.1

由实验结果可以看出，与其他 MIL 算法相比，MILICA 算法表现出了更好的分类性能，验证了 MILICA 算法在解决多示例学习问题方面的有效性。究其原因，主要是 MILICA 能提取出具有代表性的示例，并且覆盖的示例数能有效地表示这些示例的代表程度。

3.2 多类图像分类

图像分类是多示例学习最成功的应用领域之一。为衡量本文算法在多示例图像分类中的作用，本节实验选用 20 类 JPEG 格式的图像集，每类 100 幅，共计 2 000 幅，均来自 COREL 图像库。将每幅图像作为一个包，包中的示例即表征图像区域特征的显著特征点，每个特征点由一个 9 维的特征向量表示。表 4 给出了 COREL 数据集的详细信息。

本组实验分两部分进行，分别选用前 10 类共计 1 000 幅和全部 20 类共计 2 000 幅图像进行训练和测试。实验中，从每类图像集中随机选取 50%用于训练，其余的用于测试。实验采用 5 次 2-fold CV 方法进行，且每次通过调整参数 σ 的值获得不同的核心示例集，取 5 次交叉验证结果的平均值作为最终的结果。表 5 给出 MILICA 的不同 g 值对应的平均分类准确度。表 6 为 MILICA 和其他 10 种算法的平均分类准确度和 95%的置信区间的比较。其中，比较算法的结果分别取自文献[12,13,16,17,19～21,35]给出的最好结果。表 8 按照算法是否基于核心示例提取的思想分为 3 部分，其中，基于示例提取的算法对应于第 2～7 行，非示例提取算法对应于第 8～11 行，除本文算法之外的其余算法在各个数据集上的平均分类准确度的均值对应于最后一行 Average 的值。表 5 和表 6 中粗体字为表中最好的结果。

观察表 5 发现，当 $g=0.2$ 和 $g=0.8$ 时，MILICA 分别在 1 000 幅和 2 000 幅图像上得到最高的分类精度，为 86.9%和 72.6%，表 2 的准确度高于表 5 列出的其他多示例学习算法。

从表 6 可知，MILICA 算法在 2 000 幅图像上的分类准确度低于 1 000 幅。这是因为前者包含了更多的噪声示例，使得提取核心示例的过程受到干扰，从而弱化了分类预测的结果。同时，某些类别的图像存在很多视觉上的相似、语义上关联的区域，例如沙滩、山川和瀑布均含有水，也会降低分类准确度。

从实验结果还可以看出，基于示例提取的算法在两组图像数据集上的准确度均高于非示例提取算法，这说明所选取的示例对类别的正确判断具有显著作用。MILICA 在前 1 000 幅和 2 000 幅图像上的平均分类准确度均高于表 6 中其他所有算法，且比其余算法的准确度平均值分别高出 7.4%和 8.5%，

相应的95%的置信区间均小于表6中其他算法。这表明本文提出的MILICA算法不仅适用于解决两类多示例问题,在多类图像分类问题中也有效。

表4 每类图像提取得到的平均示例数

Tab.4 Average numbers of instances extracted from each kind of image

编号	图像类名称	示例数/包
0	非洲土著居民	4.84
1	沙滩	3.54
2	建筑物	3.10
3	公共汽车	7.59
4	恐龙	2.00
5	大象	3.02
6	鲜花	4.46
7	马	3.89
8	山川	3.38
9	食物	7.24
10	狗	3.80
11	蜥蜴	2.80
12	时装模特	5.19
13	夕阳	3.52
14	小轿车	4.93
15	瀑布	2.56
16	家具	2.30
17	战舰	4.32
18	滑雪	3.34
19	沙漠	3.65

表5 不同g值下MILCA在Corel图像数据集上的平均分类准确度

Tab.5 Average classification accuracy of MILIA over different values of g on the Corel datasets %

g	1 000幅图像	2 000幅图像
0	84.2	69.6
0.2	**86.9**	65.6
0.4	79.8	70.3
0.6	78.2	71.0
0.8	81.2	**72.6**
1.0	85.5	67.2

表6 不同算法在Corel图像数据集上的分类结果比较

Tab.6 Comparison of the performance of various MIL algorithms on the Corel datasets %

算法	1 000幅图像	2 000幅图像
MILICA	**86.9**:[**85.5**,**88.4**]	**72.6**:[**71.4**,**74.1**]
MKSVM-MIL	85.2:[84.1,86.3]	71.3:[70.1,72.5]
MILDS	83.0	69.4
MILD_B	79.6	67.7
MILIS	83.8:[82.5,85.1]	70.1:[68.5,71.8]
MILES	82.6:[81.4,83.7]	68.7:[67.3,70.1]
DD-SVM	81.5:[78.5,84.5]	67.5:[66.1,68.9]
mi-SVM	76.4:[75.3,77.5]	53.7:[52.2,55.2]
MI-SVM	74.7:[74.1,75.3]	54.6:[53.1,56.1]
MissSVM	78.0:[75.8,80.2]	65.2:[62.0,68.3]
Kmeans-SVM	69.8:[67.9,71.7]	52.3:[51.6,52.9]
Average	79.5	64.1

从表2和表5可看出,在多数数据集上并非当$g=0$或1.0时取得最好的效果,通过不断排除正包中的假正例,选出适当数量的具有代表性的示例有利于提高分类准确度。

4 结束语

本文从改变包的表现形式出发,结合核心示例提取方法,提出MILCA算法。该算法首先利用距离函数Hausdorff和CA构造初始核心示例集,然后使用CA算法计算这些示例的代表程度,降低假正例对分类效果的影响,并得到最终的核心示例集;最后利用核心示例集将包转变成向量的形式,从而将多示例学习转为监督单示例学习。实验结果表明,与已有的多示例学习方法相比,MILCA算法取得了较好的效果。未来的工作是研究如何更有效地选取初始示例,并对提取的核心示例集进行约简,以及尝试其他方法计算所提取示例的代表程度,并将其应用于多示例多标签学习。

参考文献:

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

[2] Li Daxiang, Wang Jing, Zhao Xiaoqiang. Multiple kernel-based multi-instance learning algorithm for image classification[J]. Journal of Visual Communication and Image Representation, 2014, 25(5):1112-1117.

[3] Zafra A, Pechenizkiy M, Ventura S. ReliefF-MI: An extension of relief to multiple instance learning[J]. Neurocomputing, 2012, 75(1):210-218.

[4] Wu Jia, Zhu Xingquan, Zhang Chenqi, et al. Multi- instance learning from positive and unlabeled bags[C]// 18th PAKDD: Advances in Knowledge Discovery and Data Mining. Switzerland: Springer International Publishing, 2014:237-248.

[5] Ding Xinmiao, Li Bing, Xiong Weihua, et al. Multi-instance multi-label learning combining hierarchical context and its application to image annotation[J]. IEEE Transactions on Multimedia, 2016, 18(8):1616-1627.

[6] Babenko B, Yang M H, Belongie S. Robust object tracking with online multiple instance learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33 (8):1619-1632.

[7] Qi Zhiquan, Xu Yitian, Wang Laisheng, et al. Online multiple-instance boosting for object detection[J]. Neurocomputing, 2011, 74 (10):1769-1775.

[8] Maron O, Lozano- Pérez T. A framework for multiple-instance learning[C]// Advances in Neutral Information Processing Systems 10 (NIPS'97). Cambridge, MA: MIT Press, 1998:570-576.

[9] Zhang Q, Goldman A S. EM-DD: An improved multi-instance learning technique[C]// Advances in Neutral Information Processing Systems 14. Cambridge, MA: MIT Press, 2002:1073-1080.

[10] Wang Xinqi, Wei Dan, Cheng Hui, et al. Multi-instance learning based on representative instance and feature mapping[J]. Neurocomputing, 2016, 216:790-796.

[11] Vanwinckelen G, Tragante do O V, Fierens D, et al. Instance-level accuracy versus bag-level accuracy in multi-instance learning[J]. Data Mining and Knowledge Discovery, 2016, 30(2):313-341.

[12] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning[C]// Advance in Neutral Information Processing System. [S. l.]:IEEE,2003:561- 568.

[13] Zhou Zhihua, Xu Junming. On the relation between multi-instance learning and semi-supervised learning[C]// Proceedings of the 24th ICML. Corvalis, Oregon:ICML, 2007:1167-1174.

[14] Shao Yuanhai, Yang Zhixia, Wang Xiaobo, et al. Multiple instance twin support vector machines[J]. ISORA, 2010(8): 433-442.

[15] Qi Zhiquan, Tian Ying, Yu Xiaodan, et al. A multi-instance learning algorithm based on nonparallel classifier[J]. Applied Mathematics and Computation, 2014,241:233-241.

[16] Chen Yixin, Wang J Z. Image categorization by learning and reasoning with regions[J]. Journal of Machine Learning Research, 2004,5(8):913-939.

[17] Chen Yixing, Bi Jin Bo, Wang J Z, et al. MILES: Multiple- instance learning via embedded instance selection[J]. IEEE Transaction Pattern Analysis and Machine Intelligence, 2006, 28(12): 1931-1947.

[18] Li Wujun, Yeung D Y. MILD: Multiple-instance learning via disambiguation[J]. IEEE Trans on Knowledge and Data Engineer, 2010, 22(1):76-89.

[19] Fu Zhouyu, Robles-Kelly A, Zhou Jun. MILIS: Multiple instance learning with instance selection[J]. IEEE Transaction Pattern Analysis and Machine Intelligence, 2011, 33(5): 958-977.

[20] Erdem A, Erdem E. Multiple-instance learning with instance selection via dominant sets[J]. Similarity-Based Pattern Recognition, 2011, 7005:171-191.

[21] Li Daxiang, Wang Jing, Zhao Xiaoqiang, et al. Multiple kernel-based multi-instance learning algorithm for image classification[J]. Journal of Visual Communication and Image Representation, 2014, 25(5):1112-1117.

[22] 张铃,张钹. M-P 神经元模型的几何意义及其应用[J]. 软件学报, 1999, 9(5):925-929.
Zhang Ling, Zhang Ba. A geometrical representation of M-P neural model and its applications[J]. Journal of Software, 1999, 9(5):925-929.

[23] Zhang Minling, Zhou Zhihua. Multi-instance clustering with applications to multi-instance prediction[J]. Applied Intelligence, 2009, 31(1):47-68.

[24] Zhang Dan, Wang Fei, Si Luo, et al. Maximum margin multiple instance clustering with its applications to image and text clustering[J]. IEEE Trans Neural Networks, 2011, 22(5):739-751.

[25] 张敏灵. 偏标记学习研究综述[J]. 数据采集预处理,2015, 3(1):77-87.
Zhang Minling. Research on partial label learning[J]. Journal of Data Acquisition and Processing, 2015, 30(1):77-87.

[26] 甘睿,印鉴. 通过挖掘示例中的概念来解决多示例学习问题[J]. 软件学报,2011,48(S2):73-78.
Gan Rui, Yin Jian. Solving multi-instance learning problem with mining concept in instances[J]. Journal of Computer Re-

search and Development, 2011, 48(S2):73-78.
[27] Zhou Zhihua, Sun Yuyin, Li Yufeng. Multi-instance learning by treating instance as Non-I. I. D samples[C]// Proceedings of the 26th International Conference on Machine Learning. New York: ACM Press, 2009:1249-1256.
[28] Zhao Shu, Rui Chen, Zhang Yanping. MICkNN: Multi-instance covering kNN algorithm[J]. Tsinghua Science and Technology, 2013, 18(4):360-368.
[29] Zhang Minglin, Huang Shengjun. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176(1):2291-2320.
[30] Hajimirsadeghi H, Mori G. Multi-Instance classification by max-margin training of cardinality-based markow networks[J]. IEEE Transactions On Pattern Analysis And Machine Intellgence,2016, 99(3):1-15.
[31] Xu Y Y. Multiple-instance learning based decision neural networks for image retrieval and classification[J]. Neurocomputing, 2016, 171(1):826-836.
[32] Li Yan, Tax D, Duin R, et al. Multiple-instance learning as a classifier combining problem[J]. Pattern Recognition, 2013, 46(3):865-874.
[33] Doran G, Ray S. Multiple-instance learning from distributions[J]. Journal of Machine Learning Research, 2016, 17(128):1-50.
[34] Liu Qiang, Zhou Sihang, Zhu Chengzhang, et al. MI-ELM: Highly efficient multi-instance learning based on hierarchical extreme learning machine[J]. Neurocomputing, 2016, 173:1044-1053.
[35] Csurka G, Bray C, Dance C, et al. Visual categorization with bags of keypoints[C]// Proc ECCV Workshop on Statistical Learning in Computer Vision. Prague, Czech:[s. n.], 2004:59-74.

作者简介:

董露露(1991-),女,讲师,研究方向:数据挖掘、机器学习等,E-mail:851601547@qq.com。

谢飞(1980-),男,博士,副教授,研究方向:人工智能、数据挖掘及文本处理。

章程(1982-),男,博士,讲师,研究方向:软件工程和数据挖掘。

(编辑:陈珺)