半监督多示例核

张 钢[1,2] 印 鉴[2] 程良伦[1] 钟钦灵[3]
(广东工业大学自动化学院 广州 510006)[1] (中山大学计算机科学系 广州 510275)[2]
(黄埔职业技术学校数学系 广州 510731)[3]

摘 要 在多示例学习中引入利用未标记示例的机制,能降低训练的成本并提高学习器的泛化能力。当前半监督多示例学习算法大部分是基于对包中的每一个示例进行标记,把多示例学习转化为一个单示例半监督学习问题。考虑到包的类标记由包中示例及包的结构决定,提出一种直接在包层次上进行半监督学习的多示例学习算法。通过定义多示例核,利用所有包(有标记和未标记)计算包层次的图拉普拉斯矩阵,作为优化目标中的光滑性惩罚项。在多示例核所张成的RKHS空间中寻找最优解被归结为确定一个经过未标记数据修改的多示例核函数,它能直接用在经典的核学习方法上。在实验数据集上对算法进行了测试,并和已有的算法进行了比较。实验结果表明,基于半监督多示例核的算法能够使用更少量的训练数据而达到与监督学习算法同样的精度,在有标记数据集相同的情况下利用未标记数据能有效地提高学习器的泛化能力。

关键词 多示例学习,半监督学习,多示例核,支持向量机
中图法分类号 TP391.41 **文献标识码** A

Semi-supervised Multi-instance Kernel

ZHANG Gang[1,2] YIN Jian[2] CHENG Liang-lun[1] ZHONG Qin-ling[3]
(Faculty of Automation, Guangdong University of Technology, Guangzhou 510006, China)[1]
(Department of Computer Science, SUN YAT-SEN University, Guangzhou 510275, China)[2]
(Department of Mathematics, Huangpu Vocational Technical School, Guangzhou 510731, China)[3]

Abstract In multi-instance learning, mechanism of making use of unlabeled instance would cut down training cost and increase generalization ability of the learner. Current algorithms perform semi-supervised multi-instance learning mainly by labeling each instance in bags and transferring multi-instance learning problem to single-instance semi-supervised ones. In this paper we introduced a bag-level semi-supervised learning framework with the idea that bag's label is determined by its instances and structure. With definition of multi-instance kernel, all bags(labeled and unlabeled) were used to calculate bag-level graph laplacian, which is a penalization term added to the optimization goal. We turned this problem into an optimization problem in RKHS and got a modified multi-instance kernel function by unlabeled data as result that can be directly used in traditional kernel learning framework. We performed experiment in ALOI and internet image datasets and compareed it with related algorithms. Experiment result shows that the proposed method can get the same accuracy as supervised counterpart with less labeled bags, and with the same labeled training data set the proposed method is of higher generalization ability.

Keywords Multi-instance learning, Semi-supervised learning, Multi-instance kernel, Support vector machine

1 概述

多示例学习始于研究者对药物活性预测问题的研究[1],并形成了一种独特的机器学习框架[2]。在多示例学习问题中,样本是包含多个示例的包,若包中至少包含一个正例,则该包为正包;若包中全部示例均为负例,则该包为负包。与传统的监督学习不同,在多示例学习问题中,包中的示例并没有标记,因此,传统的单示例学习方法并不直接适用于多示例学习问题。

自多示例学习框架提出以来,它就受到了研究者的广泛关注。提出了一系列的学习算法,这些多示例学习算法大多是对单示例学习算法进行修改,使之可以处理多示例数据。其主要目标是根据多示例学习的定义,寻找一个能够给出包中每一个示例的概念标记的假设,最终得到每个包的概念标

到稿日期:2010-10-21 返修日期:2011-02-09 本文受国家自然科学基金项目(U0935002),广东省自然科学基金项目(07117421,8351009001000002),广东工业大学高教研究基金项目(2009D06)资助。

张 钢 男,硕士,讲师,主要研究方向为机器学习、数据挖掘、知识发现,E-mail:ipx@gdut.edu.cn;印 鉴 男,博士,教授,主要研究方向为数据挖掘、机器学习;程良伦 男,博士,教授,主要研究方向为自动化装备、智能网络、传感器网络、RFID网络;钟钦灵 男,学士,教师,主要研究方向为应用数学、机器学习中的数学理论、计算机网络。

记[2]，形式上求解如下优化问题：

$$\mathrm{argmin}_h \sum_{i=1}^{l} Loss(B_i, y_i, h) + \alpha \cdot \| h \|_K^2 + \beta \cdot \| h \|_I^2 \quad (1)$$

式中，$h: R^d \to [-1,1]$；d 为每一个示例的维数；$Loss$ 表示一个损失函数，计算假设 h 作用在包 B_i 的示例中的输出和真实示例概念标记 y_i 之间的差异；$\| h \|_K^2$ 为假设在特征空间中的复杂性度量；$\| h \|_I^2$ 为假设在输入空间的复杂性度量。通过求解 h，把多示例学习转化为经典的单示例学习问题的求解。

考虑到在实际应用中包的概念标记的获取需要很大的代价，学者们研究在多示例学习框架中引入半监督学习的机制，产生了半监督多示例学习，希望通过未标记的包来提高学习器的预测精度[3,4]。文献[3]提出了一种真正的半监督多示例学习方法，其通过正则化的框架在假设空间中寻找一个假设，使之同时满足：(1)在训练集中误差最小；(2)假设在解空间中的复杂度最低(用对应 RKHS 空间的范数来表示)；(3)假设尽可能与输入空间的数据点分布一致。(3)中的数据点分布结合了训练数据和未标记数据，当未标记数据量比较大时，能够对原空间的数据分布有较好的估计。在文献[3]中，假设是在示例的层次上定义的，即以示例的属性向量表达作为输入，以在[−1,1]区间的软标签值作为输出。文献[4]提出了 MISSL 的半监督多示例学习框架，其基本思想是把半监督多示例学习问题转化为一个带有约束的半监督单示例学习问题，通过定义正包和未标记包中示例的 DD 属性[5]构造所有示例组成的图，并在图上使用标签传播算法。文献[4]的工作以一种自然的方式使用了未标记样本，算法基于半监督单示例学习中的标签传播算法，通过标签传播使未标记样本以及正包中的示例都拥有了一个以实数表示的软标签值，本质上该方法是一种面向示例的方法。

本文提出了一种基于多示例核的半监督学习方法其以文献[6,7]的工作为基础，在多示例核的定义基础上，增加光滑性惩罚项，从而有效利用未标记包。我们不寻求一个假设 f 对正包和未标记包中的所有示例进行标签，而是以某个多示例核 K 所张成的 RKHS 为假设空间，在里面寻找一个最好的以包作为输入变量的假设。形式上，本文旨在解决如下的优化问题：

$$\mathrm{argmin}_h \sum_{i=1}^{l} Loss(B_i, y_i, h) + M(B, h) \quad (2)$$

式中，$h: B \to [-1,1]$，B 为包的集合；$Loss$ 为损失函数，计算假设 h 作用在包 $Loss$ 上的输出和真实的包概念标记 y_i 之间的差异；$M(B,h)$为对样本空间光滑性的惩罚项，在单示例学习的场合，一般用图拉普拉斯来计算，在多示例学习的场合需要另外定义，在指定的核函数所张成的 RKHS 空间中，由 Representer 定理可以得到式(2)的最优解[7,14]。

本研究的动机来源于多示例核和数据依赖核，Gartner 在文献[6]中提出多示例核，其直接计算包在特征空间中的内积，使传统的核学习方法可以应用于多示例学习框架。数据依赖核是 Sindhwani 等在文献[7]中提出的，由全体数据(包括有标记和未标记)构造一个 RKHS 空间中的线性变换，修改特征空间的内积定义，使特征空间的距离函数反映数据流形，从而实现半监督学习。算法利用未标记样本提升学习器的泛化能力，对未标记样本的利用通过半监督核完成。在 RKHS 空间中的最优假设问题中引入对图光滑性的惩罚项，使优化的目标函数能够反映数据样本的光滑性。本文的贡献如下：

(1)扩展了多示例核的定义

考虑几种不同角度反映多示例包之间关系的核函数，包括顶点核、边核、改进的顶点边加权核，以及图相关的核。

(2)定义了数据依赖的多示例核

利用文献[7]中的方法对多示例核进行半监督化，利用未标记数据包在特征空间的分布修改多示例核的距离度量。

(3)通过修改的图拉普拉斯矩阵定义包之间的光滑性

以往的图拉普拉斯矩阵均是定义在单示例样本之间，本文研究多示例样本的光滑性惩罚项，定义包层次的图拉普拉斯矩阵。

本文第 2 节阐述多示例核的构造；第 3 节介绍多示例核的半监督化，这部分主要介绍利用单示例核函数的结论，在 RKHS 空间中修改原有内积定义达到多示例核的半监督化，并讨论了在半监督化过程中光滑性惩罚项的定义；第 4 节是实验数据集描述、实验结果及讨论；最后是结论。

2 多示例核

形式上定义多示例学习问题：

设 Q 为示例空间，数据集 $D=\{(X_1, y_1), (X_2, y_2), \cdots, (X_l, y_l)\}$，其中，$X_i=\{x_{i1}, x_{i2}, \cdots, x_{in_i}\} \subseteq Q$，$y_i \in Y=\{-1,+1\}$，$x_{ij}=[x_{ij1}, x_{ij2}, \cdots, x_{ijd}]^T$ 为第 i 个包中第 j 个示例的 d 个属性。若 $\exists k \in \{1,2,\cdots,n_i\}$ 使 x_{ik} 为正例，则包 X_i 为正包，即 $y_i=1$，反之 $y_i=-1$。多示例学习即通过有概念标记的包对学习器进行训练，使之能够对未标记的包给出其概念标记。

Gartner 在文献[6]中提出了多示例核的概念，通过定义包之间的核函数，可以直接把传统的核学习方法应用到多示例学习问题中。文献[6]基于集合核(Set Kernel)提出了几种多示例的构造方法。

2.1 顶点核(Node Kernel)

顶点核是在集合上定义的核函数，计算两个包之间由于示例自身的差异而产生的差异性。形式上定义点核如下：

$$K_{node}(X_i, X_j) = \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} k(x_{ip}, x_{jq}) \quad (3)$$

式中，X_i 是包，x_{ip} 是包 X_i 中的第 p 个示例，k 是一般的核函数，如 RBF 函数。顶点核的实质是不同包的元素两两的核函数值之和。

2.2 边核(Edge Kernel)

对包中的示例构建 KNN 图或 ε-Ball 图[9,11]，建立起包中示例之间的联系。

$$K_{edge}(X_i, X_j) = \sum_{p=1}^{en_i} \sum_{q=1}^{en_j} k(e_{ip}, e_{jq})$$

式中，k 是核函数，en_i 为包 X_i 中所包含的边数。沿用文献[11]的定义，设 e 是连接顶点 u 和 v 的一条边，则 e 可以表示为一个四元组，$[d_u, p_u, d_v, p_v]^T$，其中，d_u 和 d_v 分别是顶点 u 和 v 的度，p_u 是连接 u 和 v 的边的权重占连接 u 的所有边的权重的比例，p_v 类似。边权重的计算按欧氏距离的倒数或 Heat kernel 的方式计算。

2.3 改进的图核(miGraph)

文献[11]中提出了一种改进的图核 miGraph，同时考虑了顶点自身和顶点之间的关系。首先构建 KNN 图或 ε-Ball 图，定义图的离散化边权重矩阵 W^i 为：

$$W^i_{ab}=1 \text{ iff } \| x_{ia} - x_{ib} \| \leqslant \delta \quad (4)$$

$$W^i_{ab}=0 \text{ otherwise}$$

文献[11]中认为仅考虑包中的示例是否有关系，而不考

虑示例间关系的强弱，已经能够充分反映包的特征，定义改进的图核如下：

$$K_g(X_i,X_j)=\frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}W_{ia}\cdot W_{jb}\cdot k(x_{ia},x_{jb})}{\sum_{a=1}^{n_i}W_{ia}\cdot\sum_{b=1}^{n_j}W_{jb}} \tag{5}$$

式中，$W_{ia}=\frac{1}{\sum_{p=1}^{n_i}W_{ap}^{i}}$，$W_{jb}=\frac{1}{\sum_{p=1}^{n_i}W_{bp}^{i}}$，$n_i$ 为包 X_i 中元素的个数，n_j 为包 X_j 中的元素个数，k 是顶点核。该图核的计算比直接计算边核简单，且在有监督的多示例学习的场景下有较好的性能[11]。

多示例核的定义为多示例学习提供了一种核学习框架，在此框架下，把多示例样本映射到特征空间中，通过多示例核函数计算样本在特征空间中的差异性，进而直接利用现有的核学习算法（如 SVM）求解。

3 半监督多示例核（SSLMIK）

3.1 数据依赖的多示例核

现有的多示例核只能用于监督学习，现有的多示例学习的核方法也仅限于监督学习。基于单示例学习中的数据依赖核的思想，提出了数据依赖的多示例核构造方法。

考虑一个半监督多示例学习问题，形式定义如下：

设 Q 为示例空间，数据集 $D=\{(X_1,y_1),(X_2,y_2),\cdots,(X_l,y_l)\}\cup\{X_{l+1},\cdots,X_n\}$，由 l 个有标记包和 $n-l$ 个未标记包组成，$X_i=\{x_{i1},x_{i2},\cdots,x_{in_i}\}\subseteq Q$，$y_i\in Y=\{-1,+1\}$，$x_{ij}=[x_{ij1},x_{ij2},\cdots,x_{ijd}]^T$，为第 i 个包中第 j 个示例的 d 个属性。学习目标通过数据集 D 找到一个假设 h，使之对未标记包（可能在 D 中也可能不在 D 中）的概念标记有最好的预测能力。

由于在现实中有标记数据的获取通常情况下要花费很大的代价，而未标记数据较容易大量获得，因此人们希望通过未标记数据提高假设的预测能力。基于文献[7]的理论，我们认为使用未标记数据改善现有核函数的距离分布，有助于提高最终学习器的性能，而更为重要的是，半监督核可以直接嵌入一个成熟的监督核学习器。

考虑通过与数据点相关的映射修改原有核函数所张成的 RKHS 空间的内积。设 f 和 g 为多示例核 V 所张成的 RKHS 空间 H 中的两个假设，修改 H 中的内积定义，得到新的空间 H'：

$$\langle f,g\rangle_{H'}=\langle f,g\rangle_H+\langle Sf,Sg\rangle_V \tag{6}$$

H 与 H' 的差别仅在于定义了不同的内积。S 是一个映射，$S:H\to V=R^n$，$Sf=(f(x_1),\cdots,f(x_n))$，$V$ 为一个线性空间，$\|Sf\|_V^2=f^TMf$，M 为半正定矩阵，表示线性空间 V 中的一种参数化距离。文献[7,12]证明了 H' 仍为一个 RKHS，通过待定系数法可以求出 H' 的核函数的表达式：

$$k'(X_i,X_j)=k(X_i,X_j)-k_{X_i}^{t}(I+MK)^{-1}MK_{X_j} \tag{7}$$

式中，$k(X_i,X_j)$ 为一个已有的多示例核，如 K_{node} 或 K_{edge}；$k_{X_i}^{t}=(k(X_1,X_i),\cdots,k(X_n,X_i))^T$，$X_i(i=1\cdots n)$ 为训练学习器的 n 个包（包括有标记和未标记）；K 为核函数矩阵，$K_{ij}=k(X_i,X_j)$。

式(6)和式(7)表明了对于一个现有的核，可以通过一个与数据点依赖的映射对原有的核进行修改，使之能够更贴近数据的分布。

3.2 多示例样本的光滑性度量

M 的几何意义为反映样本点流形光滑程度的一个矩阵，在文献[9]中使用了样本点所构成的图拉普拉斯的整数次方，即 $M=L^p$，其中，$L=D-W$，$W_{ij}=e^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}}$（$i$ 和 j 在图中有边相连），否则 $W_{ij}=0$；$D_{ii}=\sum_i W_{ij}$，$D_{ij}=0(i\neq j)$。图拉普拉斯多用于正则化问题的求解中，作为对最终目标函数光滑程度的一种惩罚[15]。

考虑到处理的样本点为多示例数据，不能直接应用欧氏距离公式，从而无法直接计算图拉普拉斯。在多示例学习中，包中样本的欧氏距离不能直接反映包间的相似程度，因而直接从包中样本计算图拉普拉斯是不合适的。文献[3]把包“打散”，即计算所有示例的图拉普拉斯，通过半监督学习方法确定示例的标记，最终确定包的标记。文献[3,4]的方法是一种使用单示例样本的图拉普拉斯来处理多示例学习的方法，而本文的工作希望得到一个完全针对包数据的光滑性度量，我们借助度量学习中的思想，定义几种不同的包之间的度量，直接计算包层面的图拉普拉斯。为此，定义以下 3 种包间距离：

(1)包中心点距离度量（D_{Center}）

这是最简单且最为直接的计算包之间距离的方法，计算每个包的质量中心，以质量中心代表每个包计算包之间的距离。

$$D_{Center}(X_i,X_j)=\|\bar{x}_i-\bar{x}_j\|^2 \tag{8}$$

式中，$\bar{x}_i$ 为包 X_i 中所有示例的平均值。

(2)集合核距离（D_{Set}）

集合核距离实质是两个包中示例距离的平均值，即标准化后的基于欧氏距离的集合核。

$$D_{Set}(X_i,X_j)=\frac{\sum_{m,n}\|x_{im}-x_{jn}\|^2}{|x_i|\cdot|x_j|} \tag{9}$$

(3)最小示例距离（D_{Min}）

利用 Citation-KNN 的思想，计算两个包中距离最近的示例之间的距离作为两个包的距离。

$$D_{Min}(X_i,X_j)=\min_{m,n}\|x_{im}-x_{jn}^2\| \tag{10}$$

3 种距离都是基于包中示例的欧氏距离进行计算，但其中对包的相似性定义不同，从而导致其在学习问题中所得到的对于光滑性的度量不同。算法描述如图 1 所示。

半监督多示例核学习算法（SSLMIK）
Begin
　Input：有标记包 L 和无标记包 U
　通过集合 L 和 U，利用式(8)、式(9)和式(10)计算 M
　利用式(7)计算 k′
　learner=trainSVM(L,k′)//trainSVM 为标准的 SVM 学习器训练过程
　Output：elarner
End

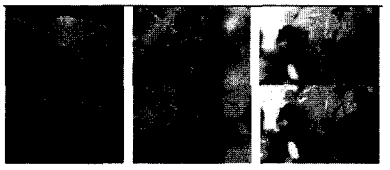
图 1　SSLMIK 算法流程图

4 实验和讨论

实验主要在图像数据集上进行，包括 ALOI（Amsterdam Library of Object Images）对象数据集[13]和网络图像数据集，将本文算法与 MISSL[4] 进行比较，并以 MI-Kernel[6] 方法的准确率作为算法的有效性基线。

4.1 图像数据处理

参考文献[12]的方法，把图片转化为多示例学习中的包，转化的方法为先对图片进行 LUV 颜色空间的转换，然后进行小波变换，对每个点的小波系数进行聚类，得到若干个区域，将整幅图像看作一个包，把区域中心点的属性值作为包中的每一个示例，在本算法中，以九元组表示一个示例。图 2 是图像区域划分之后的示意图。



第 1 行：原始图像 第 2 行：经过算法处理后的图像

图 2　图像处理效果示意图

在算法处理的过程中使用 KMEANS 算法进行聚类，而标准的 KMEANS 算法必须事先指出聚类的数目。若使用较复杂的聚类算法，如 Global-KMEANS 则可以自动寻找在某种最优指标下的聚类数目，但计算代价比 KMEANS 大得多。为了提高图像处理的速度，本研究采用了标准的 KMEANS，预先根据经验设置一个比合理聚类数目稍偏大的数目，然后在 KMEANS 算法处理完成后把包含点数小于一定阈值的聚类删去。算法的实际运行表明这样的处理是比较合理的，既不会因为细小聚类影响算法的运行，也不会为了追求精确的最优聚类数目而大大增加了计算时间。

4.2 实验数据集

实验数据集为 ALOI 和网络图片数据集。ALOI 是一个对象图片数据集，包含 1000 个对象，其中每个对象的图像数据中包含了不同光照、不同角度的若干幅图像，每幅图像的大小是宽 192、高 144 像素的图片的 24bit RGB 彩色图片。采用了每个对象 72 种不同角度的图像作为实验数据。

网络图像数据集来源于 Internet，即通过 Internet 收集的属于不同的主题的图像。在本实验中通过 Internet 收集 4 个主题(flower、shoes、boots 和 people)的图像进行算法的测试。数据集中的图像大小不统一，均为彩色图像，通过 4.1 节的处理方法进行自适应，得到统一的向量表示形式。表 1 描述了 Internet 图像主题数据集。

表 1　Internet 图像主题数据集

名称	图像数量	大小	说明
flower	140	<500k	主体单一
shoe	120	<500k	主体较单一
boots	150	<500k	混合
people	20	<500k	大场景

在测试图片数据集上设置若干场景测试算法对图片主题的识别能力，包括对同一对象的识别能力以及同类对象的识别能力。同时选取与正包属同一大类和不同大类的负包进行学习器训练。

4.3 实验结果及讨论

4.3.1 ALOI 数据集

进行 20 次随机测试，在每一次测试中，从 1000 个对象中随机选取 l 个作为正例。每个对象中包含不同角度的 72 幅图像，从中取出 n 幅作为正训练样例，再从余下的对象中随机抽取角度随机的 n 幅作为负训练样例。测试样例中的正例从被选对象余下的图像中选取，负例从其余对象中随机选取。作为对比，同时在相同的数据集上运行 MISSL[4] 算法，并以 MI-Kernel[6] 算法的分类正确率作为基线。SSLMIK 中采用 K_{node} 作为半监督核的核函数。

图 3 中，X 轴代表训练样本所占全体样本的比例，每个刻度代表 10%，Y 轴代表分类的正确率。正例从单一对象的图像中随机抽取，负例从余下对象的图像中抽取。其中 SSLMIK 是本文的方法。对于 MI-Kernel，由于它不是半监督学习算法，只使用有标记的训练样本进行训练。对于 MISSL 和 SSLMIK，采用转导学习的方法，即把训练和测试样本作为学习器的输入进行训练。从图 3 中可以看出，SSLMIK 在有标记训练样本规模较小的情况下能够得到较好的分类正确率，这是由于 ALOI 中相同对象的图像之间有一定的连续性，如图 4 所示。SSLMIK 的光滑性度量能够很好地从未标记数据中学到这种连续性。

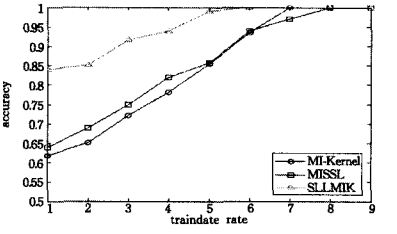


图 3　ALOI 分类正确率

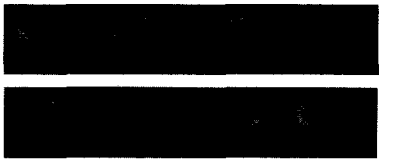


图 4　ALOI 数据集中对象不同角度图像

为了进一步说明 SSLMIK 的有效性，我们对训练数据集进行人工选择，从每个对象中选出 4 个最有代表性的图像作为训练数据，分别是正面、背面和两个侧面，如图 5 所示。

图 5　有代表性的 4 个角度的图像

随机选取 20 个对象，使用 SSLMIK 和 MI-Kernel 进行测试，SSLMIK 采用转导学习的方法，而 MI-Kernel 直接使用 4 幅图像作为训练集，分类的正确率及其方差如表 2 所列。

表 2　人工选择角度训练集的分类正确率

SSLMIK	MI-Kernel
98.3%±0.84%	64.7%±1.48%

4.3.2 Internet 图像主题数据集

用 K_{edge} 作为半监督核的核函数，M 的计算采用 D_{Set}，在半监督学习的场景中，随机选取正负包的 40%作为有标记数据，其余作为未标记数据。作为对比，在相同的数据集中运行 MI-Kernel 测试。K_{edge} 的计算选用 5-NN 图，采用欧氏距离的倒数作为 5-NN 图的边权重。结果如表 3 所列。

表 3　Internet 主题数据集测试结果

类别	SSLMIK	MI-Kernel
boots vs. flower	73.1%±0.21%	76.2%±0.34%
boots vs. shoe	70.9%±0.48%	62.8%±0.52%
flower vs. shoe	80.2%±0.27%	75.4%±0.3%
flower vs. people	76.6%±0.4%	78.1%±0.24%
boots vs. people	63.2%±0.32%	61.5%±0.61%

(下转第 236 页)

$bn_C(Y)=\{x_1,x_2,x_3,x_4,x_5,x_7,x_8\}$

经过计算得知

$h_Y(x_1)=h_Y(x_8)=\{x_1,x_8\}$

$h_Y(x_3)=\phi \quad h_Y(x_4)=\phi$

$h_Y(x_2)=h_Y(x_5)=h_Y(x_7)=\{x_2,x_5,x_7\}$

$l_X(x_1)=\{x_1,x_4\},l_X(x_2)=\{x_4\}$

$l_X(x_3)=\phi \quad l_X(x_4)=\phi$

$l_Y(x_5)=\{x_4\} \quad l_Y(x_7)=\{x_4\}$

$l_Y(x_8)=\{x_4\}$

由引理 2 知，$\overline{Z}_X(Y)=\{x_3\}$，从而得知

$\overline{C}(X\cap Y)=\overline{C}(X)\cap\overline{C}(Y)-\overline{Z}_X(Y)$。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11: 341-356

[2] Pawlak Z. Rough sets: Theoretical Aspects of Reasoning about Data[J]. Kluwer Academic Publishers, Boston, 1991

[3] Zhang Hua-guang, Liang Hong-li, et al. Two new operators in rough set theory with applications to fuzzy sets[J]. Information Sciences, 2004, 166: 147-165

[4] 徐忠印，王勤. 覆盖粗糙集模型的性质[J]. 河南师范大学学报：自然科学版，2005，33(1)：130-132

[5] 杨勇，朱晓钟，李廉. 覆盖粗糙集的公理化[J]. 计算机科学，2009，36(5)：181-182

[6] 刘瑞新，孙士保，秦克云. 变精度覆盖粗糙集[J]. 计算机工程与应用，2008，44(12)：47-50

[7] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001

[8] 苗夺谦，王国胤，等. 粒计算：过去、现在与展望[M]. 北京：科学出版社，2007

(上接第 223 页)

SSLMIK 的分类正确率略好于 MI-Kernel，我们认为这是由于数据集中的数据的分布规律性并不突出，导致半监督学习的效果不明显，特别是 boots 数据集的图像中同时包含主题(boots)和非主题(如模特、背景、装饰物等)都会对算法的分类正确率有较大的影响，且同一主题的图像之间的分布未能呈现明显的规律性，因此本文的算法所计算的矩阵 M 不能很好地反映真实的数据分布。

结束语 本文讨论了利用多示例学习算法解决图像主题识别的问题，其利用多示例学习中的核学习方法，把每一个图像看作多示例学习框架中的包，进而定义包之间的核函数，把包看作一个 KNN 图或 ε-Ball 图，由图核的定义得出顶点核和边核，并利用图的光滑性理论推导出半监督多示例核，能够利用未标记包所蕴含的数据分布规律改善学习器的性能。这是一个把半监督核应用到多示例学习问题的尝试，能够有效解决标记样本获取代价大的问题，并能通过已有未标记样本最大限度提升学习器的泛化能力，是一种完全的半监督学习方法。

在真实的数码照片数据集和 ALOI 对象数据集上进行算法的性能测试，实验结果表明本文算法对于图像的主题识别问题是有效的，相比以往的算法有较高的正确率和泛化能力，且引入未标记的示例能有效提高学习器的性能和泛化能力，可以推广到图像主题识别的其它方面。

算法的一个核心问题是如何在所定义的边核和点核中结合反映数据流形光滑程度的度量，我们采用的方法是使用图的拉普拉斯矩阵，但由于处理的是多示例学习问题，每个包中有一定数量的实例，而本文的拉普拉斯的计算方法是针对单示例学习问题的，使用包中心作为代表示例的可行性也值得研究，这将是后续研究工作的重点之一。

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] Zhou Zhi-hua. Multi-Instance Learning: A Survey[R]. CS, Nanjing University, 2004

[3] Jia Yang-qing, Zhang Chang-shui. Instance-level Semisupervised Multiple Instance Learning[C]// Proceedings of the 23rd Intl. Conf. AAAI, 2008: 640-645

[4] Rahmani R, Goldman S A. MISSL: Multiple-instance semi-supervised learning[C]// Proceedings of the 23rd Intl. Conf. ICML, 2006: 705-712

[5] Maron O, Lozano-Pérez T. A framework for multiple-instance learning[J]. Neural Information Processing Systems, 1998

[6] Gartner T, Flach P A, Kowalczyk A, et al. Multi-instance kernels[C]// Proceedings of the 19th Intl. Conf. ICML, 2002: 179-186

[7] Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: from transductive to semi-supervised learning[C]// Proceedings of the 22nd Intl. Conf. ICML, 2005: 824-831

[8] Zhou Zhi-hua, Sun Yu-yin, Li Yu-feng. Multi-instance learning by treating instances as non-I. I. D. samples[C]// Proceedings of the 26th Intl. Conf. ICML, 2009: 1249-1256

[9] Zhu X, Ghahramani Z, Lafferty J. Semisupervised learning using gaussian fields and harmonic functions[C]// Proceedings of the 20th Intl. Conf. ICML, 2003

[10] Zhou Zhi-hua, Xu Jun-ming. On the relation between multi-instance learning and semi-supervised learning[C]// Proceedings of the 24th Intl. Conf. ICML, 2007: 1167-1174

[11] Niyogi P. Manifold regularization and semi-supervised learning: Some theoretical analysis[R]. Department of Computer Science, University of Chicago, 2008

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

[13] Geusebroek J M, Burghouts G J, Smeulders A W M. The Amsterdam library of object images[J]. International. Journal of Computer Vision, 2005, 61: 103-112

[14] Sindhwani V, Rosenberg D S. An RKHS for multi-view learning and manifold co-regularization[C]// Proceedings of the 25th Intl. Conf. ICML, 2008

[15] 刘叶青，刘三阳，谷明涛. 一种多项式光滑的半监督支持向量机分类算法[J]. 计算机科学，2009，36(7)：179-181