

doi: 10.12052/gdutxb.180036

基于抗噪声的多任务多示例学习算法研究

黎启祥¹, 肖燕珊¹, 郝志峰², 阮奕邦¹

(1. 广东工业大学 计算机学院, 广东 广州 510006; 2. 佛山科学技术学院 数学与大数据学院, 广东 佛山 528000)

摘要: 在多示例学习中, 当训练样本数量不充足或者训练样本中存在噪声信息时, 分类器的分类性能将降低. 针对该问题, 本文提出了一种基于抗噪声的多任务多示例学习算法. 一方面, 针对训练样本中可能存在的噪声问题, 该算法赋予包中示例不同的权值, 通过迭代更新权值来降低噪声数据对预测结果的影响. 另一方面, 针对训练样本数量不充足问题, 该算法运用多任务学习策略, 通过同时训练多个学习任务, 利用任务间的关联性来提高各个分类任务的预测性能. 实验结果证明, 与现有的分类算法相比, 该方法在相同的实验条件下具有更优秀的性能.

关键词: 多示例学习; 抗噪声; 多任务学习; 关联性; 分类器

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1007-7162(2018)03-0047-07

An Algorithm Based on Multi-task Multi-instance Anti-noise Learning

Li Qi-xiang¹, Xiao Yan-shan¹, Hao Zhi-feng², Ruan Yi-bang¹

(1. School of Computers, Guangdong University of Technology, Guangzhou 510006, China; 2. School of Mathematics and Big Data, Foshan University, Foshan 528000, China)

Abstract: In multi-instance learning, classification performance may be limited due to the noisy data or a scarce amount of labeled data. To solve this problem, an algorithm based on multi-task multi-instance anti-noise learning is proposed. On the one hand, in view of the noisy data, the algorithm trains a classifier by assigning the instances in bags with different weights. And the weights of instances are updated by adopting an iterative optimization framework which decreases the influence of the noisy data. On the other hand, in view of insufficient labeled data, the classifier is extended to multi-task learning to train multiple learning tasks at the same time, so that the performance of each learning task can be improved by sharing the classification information among the tasks. Extensive experiments have showed that the proposed classification framework outperforms the existing classification methods.

Key words: multi-instance learning; anti-noise; multi-task learning; correlation; classification

多示例学习是一种在处理训练样本时, 将训练样本当成多示例包 (多个示例的集合) 来进行处理的有监督分类学习方法^[1]. 在多示例学习中, 训练集由多个多示例包构成, 每个包中包含多个单示例. 对于分类任务的训练集, 一个多示例包对应一个样本, 样本标签只和包有关, 而包中的每个示例的所属标签是无法直接获取的^[2]. 传统二分类任务中, 若多示例包中至少含有一个示例为正类示例, 则该多示例包被标记为正包; 相反, 若多示例包中的所有示例均

为负类示例, 则该多示例包被标记为负包. 对于二分类多示例问题, 分类器将会在已知标签的多示例包中进行训练, 从而对未知标签的包进行分类, 预测其所属分类为正类还是负类. 多示例学习已经取得了广泛的研究与应用, 如文档分类^[3]、图像识别^[4]、情感分析^[5-6]、人脸识别^[7]、对象检测^[8]等.

多示例学习问题提出后, 研究者们提出了一系列多示例学习算法. Andrews等^[9]在支持向量机基础上, 提出了两种基于支持向量机的多示例学习算法

收稿日期: 2018-03-05
基金项目: 国家自然科学基金资助项目(61472090, 61672169, 61472089)
作者简介: 黎启祥(1992-), 男, 硕士研究生, 研究方向为机器学习、数据挖掘、图像处理.
通信作者: 肖燕珊(1981-), 女, 教授, 博士, 主要研究方向为机器学习、图像处理与模式识别. E-mail: xiaoyanshan@189.cn

MI-SVM和mi-SVM. Qi Zhang等^[10]结合最大期望(EM)算法和多样性密度(DD)算法,提出了EM-DD算法. Jun Wang等^[11]在K-近邻(KNN)算法基础上,使用基于包层级的豪斯多夫距离来取代欧氏距离,提出了基于最小化豪斯多夫距离(minimum Hausdorff distance)的Citation-KNN算法.

这些多示例学习算法能够获得不错的分类效果的前提条件是,数据集中具有数量充足的训练样本.当数据集的训练样本数量充足时,这些多示例学习算法能够获得充分的分类信息来建立准确的预测模型.但是,当数据集的训练样本数量不充足时,由于分类信息的缺乏,训练所得模型的分类效果可能不佳^[12].针对训练集中样本数量可能不充足的情况,本文提出了基于抗噪声的多任务多示例学习算法(Multi-task Multi-instance Anti-noise Learning, MtMiAnL).一方面,本文所提出的MtMiAnL算法能够在训练样本数量不足时,通过同时训练多个学习任务,利用任务间的关联性来提高各个分类任务的预测性能.另一方面,在实际的应用中,数据可能在传输过程中受到的干扰,存在噪声信息. MtMiAnL算法通过示例加权方式,降低噪声数据对预测结果的影响.

本文提出了基于抗噪声的多任务多示例学习算法(MtMiAnL),跟现有的多示例学习算法相比,本文的主要贡献在于:(1)算法充分考虑了数据可能存在噪声信息,在训练过程中对示例进行加权,通过迭代更新权值来降低噪声对结果的影响;(2)提出了在多任务框架下的抗噪声多示例学习算法,解决了在训练样本数量不充足的情况下,通过联合多个相关任务一起学习,提高分类的精度.

1 MtMiAnL算法原理

MtMiAnL算法分为多示例加权和目标函数多任务化两部分.在多示例加权阶段,首先对样本中的示例添加初始权值,然后在训练过程中对示例的权值进行更新,以降低噪声示例对预测结果的影响.在目标函数多任务化阶段,把算法从单任务学习环境扩展到多任务学习环境,最大化利用多个任务的关联性来提高算法性能,从而解决了分类模型因样本数量不充足而造成性能不佳的问题.

2 基于抗噪声的多任务多示例学习算法

2.1 多示例加权

假设存在 T 个相似的多示例加权学习任务,则第

$t(1 < t \leq T)$ 个任务的数据集表示为 $\mathbf{D}_t$, $\mathbf{D}_t$ 由正包集合 $\mathbf{P}_t$ 和负包集合 $\mathbf{N}_t$ 组成,即 $\mathbf{D}_t = \mathbf{P}_t \cup \mathbf{N}_t$. 其中,正包集合 $\mathbf{P}_t = [\mathbf{B}_{1t}, \dots, \mathbf{B}_{it}, \dots, \mathbf{B}_{|P_t|t}]$, 负包集合 $\mathbf{N}_t = [\mathbf{B}_{1t}, \dots, \mathbf{B}_{it}, \dots, \mathbf{B}_{|N_t|t}]$, 式中 $|P_t|$ 和 $|N_t|$ 分别表示任务 t 的正包集合 $\mathbf{P}_t$ 和负包集合 $\mathbf{N}_t$ 所包含的多示例包个数. 分类标签 $Y_{it} \in \{-1, 1\}$ 为任务 t 的第 i 个多示例包 $\mathbf{B}_{it}$ 的标签. 当 $Y_{it}=1$ 时,即 $\mathbf{B}_{it} \in \mathbf{P}_t$, 多示例包 $\mathbf{B}_{it}$ 被认定为正包;相反,当 $Y_{it}=-1$ 时,即 $\mathbf{B}_{it} \in \mathbf{N}_t$, 多示例包 $\mathbf{B}_{it}$ 被认定为负包.

多示例包 $\mathbf{B}_{it}$ 中的 m 个示例 $[\mathbf{B}_{it}^1, \dots, \mathbf{B}_{it}^j, \dots, \mathbf{B}_{it}^m]$ 被赋予不同的权值 $[Q_{it}^{1(k)}, \dots, Q_{it}^{j(k)}, \dots, Q_{it}^{m(k)}]$, 权值的表达形式为

$$Q_{it}^{j(k)} = Q_{it}^{j(k-1)} e^{(Y_{it} \cdot y_{it}^j - 1) / \alpha^2}. \quad (1)$$

式(1)中, $Q_{it}^{j(k)}$ 表示包中示例 $\mathbf{B}_{it}^j$ 在第 k 次迭代的权值, $Q_{it}^{j(k-1)}$ 表示第 $k-1$ 次迭代的权值, α 为常数调节因子,一般设为1, Y_{it} 为多示例包 $\mathbf{B}_{it}$ 的标签, y_{it}^j 为包中示例 $\mathbf{B}_{it}^j$ 的标签. Y_{it} 是已知的,而 y_{it}^j 是未知的.在第 k 次迭代求解中,对于多示例包 $\mathbf{B}_{it}$,若包中示例 $\mathbf{B}_{it}^j$ 的标签 y_{it}^j 与包标签 Y_{it} 一致,则认为该示例不是噪声示例,此时 $Q_{it}^{j(k)} = Q_{it}^{j(k-1)} e^0$,示例的第 k 次迭代权值与第 $k-1$ 次权值相等,不会发生变化;相反,当示例 $\mathbf{B}_{it}^j$ 的标签 y_{it}^j 与包标签 Y_{it} 不一致时,则该示例为包中的噪声示例,故第 k 次迭代的权值为 $Q_{it}^{j(k)} = Q_{it}^{j(k-1)} e^{-2/\alpha^2}$,与第 $k-1$ 次的迭代权值相比,将适当缩小,从而降低该示例对结果的影响.

当多示例包 $\mathbf{B}_{it}$ 为正包时,用符号 p_{it} 来表示加权后的包 $\mathbf{B}_{it}$;当多示例包 $\mathbf{B}_{it}$ 为负包时,用符号 n_{it} 来表示加权后的包 $\mathbf{B}_{it}$,具体描述如下

$$p_{it} = \sum_{j=1}^{|\mathbf{B}_{it}|} Q_{it}^{j(k)} \mathbf{B}_{it}^j \Big/ \sum_{j=1}^{|\mathbf{B}_{it}|} Q_{it}^{j(k)}, \text{ if } Y_{it} = 1. \quad (2)$$

$$n_{it} = \sum_{j=1}^{|\mathbf{B}_{it}|} Q_{it}^{j(k)} \mathbf{B}_{it}^j \Big/ \sum_{j=1}^{|\mathbf{B}_{it}|} Q_{it}^{j(k)}, \text{ if } Y_{it} = -1. \quad (3)$$

式(2)和(3)中, $|\mathbf{B}_{it}|$ 表示多示例包 $\mathbf{B}_{it}$ 的示例个数.在迭代过程中, p_{it} 和 n_{it} 分别按照式(1)进行更新,能有效降低正负示例包中噪声示例对模型的影响,从而提高模型的预测精度.

根据式(2)和式(3)中正负多示例包的加权表达形式以及约束条件,可以为这 T 个多示例加权分类任务构造一个优化问题,该优化问题可以表达为

$$\begin{cases} \text{minimize:} \\ J(\mathbf{w}_t, b_t, \zeta_{it}^+, \zeta_{it}^-) = \\ \frac{1}{2} \sum_{i=1}^T \|\mathbf{w}_t\|^2 + C_0 \sum_{i=1}^T \sum_{i=1}^{|P_t|} \zeta_{it}^+ + C_0 \sum_{i=1}^T \sum_{i=1}^{|N_t|} \zeta_{it}^- . \\ \text{subject to:} \\ \mathbf{w}_t \mathbf{p}_{it} + b_t \geq G_{it} - \zeta_{it}^+, 1 \leq i \leq |P_t|, \\ \mathbf{w}_t \mathbf{n}_{it} + b_t \leq -1 + \zeta_{it}^-, 1 \leq i \leq |N_t|, \\ \zeta_{it}^+ \geq 0, \zeta_{it}^- \geq 0, 1 \leq t \leq T. \end{cases} \quad (4)$$

式(4)中, $\mathbf{w}_t$ 为分类任务 t 的向量, 约束条件中的控制变量 $G_{it} = (2 - |\mathbf{B}_{it}|) / |\mathbf{B}_{it}|$, $|\mathbf{B}_{it}|$ 表示任务 t 的正包 $\mathbf{B}_{it}$ 中的示例个数, 该控制变量 G_{it} 在约束条件中可保证包中的正示例个数至少为 1. $C_0 \geq 0$ 是惩罚系数, $\zeta_{it}^+ \geq 0$ 和 $\zeta_{it}^- \geq 0$ 是松弛变量, 使优化问题可以容纳一定的错误分类结果.

通过迭代求解最小化问题(4), 可以在保持包中正常示例的权值不变的情况下, 将包中噪声示例的权值不断减小, 从而降低噪声示例对分类器的影响.

2.2 多任务目标方程

然而, 优化问题(4)只是单纯将多个分类任务进行累加求和并求解, 并未考虑到多个任务间的关联性, 本质上还是一种单任务学习问题. 在样本数量不充足的情况下, 并不能够保证分类任务 t 的分类模型的预测性能. 因此, 下文对优化问题(4)进行了扩展, 提取任务间的关联性, 把算法扩展到多任务环境中.

在多任务学习环境中, 为了提取 T 个任务间的相似性来提高每个分类任务的预测性能, 本文将任务 t 的向量 $\mathbf{w}_t$ 表示为多任务形式: $\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t$, 向量 $\mathbf{v}_t$ 表示任务 t 所独有的向量, 向量 $\mathbf{w}_0$ 为 T 个任务所共享的向量^[13]. 因此, 引入向量的多任务形式 $\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t$ 到优化问题(4)中, 将其扩展到多任务学习环境下, 从而提出了基于抗噪声的多任务多示例学习算法模型

$$\begin{cases} \text{minimize:} \\ J(\mathbf{w}_0, \mathbf{v}_t, \zeta_{it}^+, \zeta_{it}^-) = \\ T\lambda_0 \|\mathbf{w}_0\|^2 + \lambda_1 \sum_{i=1}^T \|\mathbf{v}_i\|^2 + C_0 \sum_{i=1}^T \sum_{i=1}^{|P_t|} \zeta_{it}^+ + C_0 \sum_{i=1}^T \sum_{i=1}^{|N_t|} \zeta_{it}^- . \\ \text{subject to:} \\ (\mathbf{w}_0 + \mathbf{v}_t)^T \mathbf{p}_{it} + b_t \geq G_{it} - \zeta_{it}^+, 1 \leq i \leq |P_t|, \\ (\mathbf{w}_0 + \mathbf{v}_t)^T \mathbf{n}_{it} + b_t \leq -1 + \zeta_{it}^-, 1 \leq i \leq |N_t|, \\ \zeta_{it}^+ \geq 0, \zeta_{it}^- \geq 0, 1 \leq t \leq T. \end{cases} \quad (5)$$

式(5)中, $\lambda_0 \geq 0$ 和 $\lambda_1 \geq 0$ 为多任务权重系数, 用来控制任务之间的相似度.

当 λ_0 为一个较小的固定值, 而 λ_1 趋近于无穷大时, 训练出来的各个任务所独有的向量 $\mathbf{v}_t$ 趋近于零向

量, 则有 $\mathbf{w}_t = \mathbf{w}_0$, 所有任务的向量 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T$ 将会类似, 此时所有任务的相似性最高.

当 λ_1 为一个较小的固定值, 而 λ_0 趋近于无穷大时, 此时训练出来的任务间共享的向量 $\mathbf{w}_0$ 趋近于零向量, 则有 $\mathbf{w}_t = \mathbf{v}_t$, 所有任务的向量 $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T$ 将相互独立, 目标函数(5)退化为目标函数(4).

通过求解问题(5), 可以得到优化后的 T 个任务间所共享的向量 $\mathbf{w}_0$, 以及各个任务所独有的向量 $\mathbf{v}_t$, 此时任务 t 的分类函数为

$$F^t(\mathbf{B}_t) = \text{sign} \max_{\mathbf{B}_t^j \in \mathbf{B}_t} ((\mathbf{w}_0 + \mathbf{v}_t) \mathbf{B}_t^j + b_t). \quad (6)$$

式(6)中, $\mathbf{B}_t$ 为分类任务 t 中的未知标签多示例包, $\mathbf{B}_t^j$ 为该未知多示例包中的示例. 当 $\mathbf{B}_t$ 中的示例 $\mathbf{B}_t^j$ 为负示例时, 则有 $(\mathbf{w}_t + \mathbf{v}_t) \mathbf{B}_t^j + b_t < 0$; 相反, 当 $\mathbf{B}_t^j$ 为正示例时, 则有 $(\mathbf{w}_t + \mathbf{v}_t) \mathbf{B}_t^j + b_t > 0$. 因此, 当 $\mathbf{B}_t$ 中的所有示例都为负示例时, 根据分类函数(6)可计算得出: $F^t(\mathbf{B}_t) = -1$, 即多示例包 $\mathbf{B}_t$ 被划分为负包; 当 $\mathbf{B}_t$ 有一个或一个以上的示例为正示例时, 根据分类函数(6)可计算得出: $F^t(\mathbf{B}_t) = 1$, 即多示例包 $\mathbf{B}_t$ 被划分为正包.

2.3 优化问题求解

本文拟将式(5)转化为标准SVM形式的二次规划问题, 从而求解转化后的二次规划问题. 首先, 定义多任务分类模型的耦合参数 $\mu = T\lambda_0/\lambda_1$ 和惩罚系数 $C = C_0/\lambda_1$; 然后, 将 $|P_t|$ 个正包 $\mathbf{p}_{it}$ 与 $|N_t|$ 个负包 $\mathbf{n}_{it}$ 合并为新向量 $\mathbf{A}_{it}$, 见式(7), 将 $|P_t|$ 个变量 G_{it} 与 $|N_t|$ 个常数 1 合并为新向量 $\mathbf{I}_{it}$, 见式(8).

$$\mathbf{A}_{it} = [\mathbf{p}_{1t}, \dots, \mathbf{p}_{|P_t|t}, \mathbf{n}_{(|P_t|+1)t}, \dots, \mathbf{n}_{(|P_t|+|N_t|)t}], \quad (7)$$

$$\mathbf{I}_{it} = [G_{1t}, \dots, G_{|P_t|t}, 1, \dots, 1]. \quad (8)$$

基于上述变量 $\mathbf{A}_{it}$ 和 $\mathbf{I}_{it}$ 的定义, 问题(5)可以简化表示为

$$\begin{cases} \text{minimize:} \\ J(\mathbf{w}_0, \mathbf{v}_t, \zeta_{it}) = \\ \frac{1}{2} (\mu \|\mathbf{w}_0\|^2 + \sum_{i=1}^T \|\mathbf{v}_i\|^2) + C \sum_{i=1}^T \sum_{i=1}^{|P_t|+|N_t|} \zeta_{it} . \\ \text{subject to:} \\ Y_{it} [(\mathbf{w}_0 + \mathbf{v}_t)^T \mathbf{A}_{it} + b_t] \geq \mathbf{I}_{it} - \zeta_{it}, \\ 1 \leq i \leq |P_t| + |N_t|, \\ \zeta_{it} \geq 0, 1 \leq t \leq T. \end{cases} \quad (9)$$

为了解决问题(9), 构建两个向量

$$\mathbf{W} = [\sqrt{\mu} \mathbf{w}_0^T, \mathbf{v}_1^T, \dots, \mathbf{v}_T^T, \dots, \mathbf{v}_T^T]^T, \quad (10)$$

$$\mathbf{X}_{it} = \left[\frac{1}{\sqrt{\mu}} \mathbf{A}_{it}^T, \underbrace{0, \dots, 0}_{t-1}, \mathbf{A}_{it}^T, \underbrace{0, \dots, 0}_{T-t} \right]^T. \quad (11)$$

根据上述定义的向量,可以发现 $\|\mathbf{W}\|^2 = \mu \|\mathbf{w}_0\|^2 + \sum_{t=1}^T \|\mathbf{v}_t\|^2$ 和 $\mathbf{W}^T \mathbf{X}_{it} = (\mathbf{w}_0 + \mathbf{v}_t)^T \mathbf{A}_{it}$. 因此,问题(9)可以进一步改写为

$$\begin{cases} \text{minimize: } \frac{1}{2} \|\mathbf{W}\|^2 + C \sum_{t=1}^T \sum_{i=1}^{|\mathbf{P}_t|+|\mathbf{N}_t|} \zeta_{it}. \\ \text{subject to:} \\ Y_{it}(\mathbf{W}^T \mathbf{X}_{it} + b_t) \geq I_{it} - \zeta_{it}, \\ 1 \leq i \leq |\mathbf{P}_t| + |\mathbf{N}_t|, \\ \zeta_{it} \geq 0, 1 \leq t \leq T. \end{cases} \quad (12)$$

通过标准SVM的二次规划求解法,如LibSVM^[14],可以对问题(12)进行求解,从中得到 T 个任务间所共享的向量 $\mathbf{w}_0$,以及各个任务所独有的向量 $\mathbf{v}_t$.

2.4 算法步骤

本文算法流程如下:

输入: T 个任务数据 $\mathbf{D}_1, \dots, \mathbf{D}_t, \dots, \mathbf{D}_T$, 权值调节因子 α , 多任务分类模型的耦合参数 μ , 惩罚系数 C , 终止迭代次数 K ;

输出: T 个任务的分类器: $F^1, \dots, F^t, \dots, F^T$.

算法步骤:

1) 在任务 t 中, 初始化正包中所有示例的标签为1, 负包中所有示例的标签为-1, 设置示例 $\mathbf{B}_{it}^j$ 的初始权值为 $Q_{it}^{(0)} = 1/|\mathbf{B}_{it}|$;

2) 更新 T 个分类任务中, 正包和负包所包含的示例 $\mathbf{B}_{it}^j$ 的权值 $Q_{it}^{(k)}$;

3) 根据表达式(2)和(3), 构建正包 $\mathbf{p}_{it}$ 和负包 $\mathbf{n}_{it}$;

4) 根据定义描述式(7), (8)和(11), 构建变量 $\mathbf{A}_{it}$, $\mathbf{I}_{it}$ 和 $\mathbf{X}_{it}$;

5) 求解问题(12), 得到 T 个任务所共享的向量 $\mathbf{w}_0$ 和各个任务所独有的向量 $\mathbf{v}_t$, 并获得 T 个任务的分类器 $F^1, \dots, F^t, \dots, F^T$

6) 计算目标函数(9)的值 J_k , 若 $(J_{(k-1)} - J_k)/J_k < 0.001$ 或迭代次数 k 达到 K 时, 跳转到步骤9), 否则进入下一步;

7) 基于分类函数(6), 更新 T 个任务中所有多示例包所包含的示例的标签 y_{it}^j ;

8) $k=k+1$ 并跳转到步骤2);

9) 迭代结束, 得到每个任务的分类器.

其中, $J_{(k-1)}$ 表示第 $k-1$ 次迭代时的多任务优化问题(9)的函数值, J_k 表示在第 k 次迭代时的优化问题

(9)的函数值.

3 实验与结果分析

3.1 数据集

1) Corel数据集.

参考文献^[15]的实验设定, 在Corel数据集中选取“非洲生活”“美食”“汽车”“花”“室内”“室外”“风景”和“非风景”8个类, 组成4个任务, 即任务1、2、3和4. 每一个任务是一个二分类问题, “非洲生活”“汽车”“风景”和“室内”分别为任务1、2、3和4的正类, 而“美食”“花”“非风景”和“室外”分别为任务1、2、3和4的负类. 对于一个任务, 正类和负类各有100幅图像. 在训练过程中, 抽取20幅正类图和20幅负类图, 其余160幅图作为测试集.

2) CIFAR数据集.

选取CIFAR数据集的“卡车”“狗”“小汽车”“猫”“飞机”“鸟”“轮船”和“马”8个类, 组成四个任务, 即任务1、2、3和4. 每一个任务是一个二分类问题, “卡车”“小汽车”“飞机”和“轮船”分别为任务1、2、3和4的正类, 而“狗”“猫”“鸟”和“马”分别为任务1、2、3和4的负类. 对于一个任务, 正类和负类各有100幅图像. 在训练过程中, 抽取20幅正类图和20幅负类图, 其余160幅图作为测试集.

为了研究MtMiAnL算法在处理噪声数据方面的性能, 按照文献^[16]的实验方法, 在Corel和CIFAR数据集中, 随机选取10%的多示例包添加噪声. 具体来说, 在示例的每维数据随机添加 $[0, \sigma]$ 的噪声, 其中 σ 是示例 n 维数据的标准差.

3.2 比较算法

为了使实验更具针对性, 选择多种算法进行对比实验. 在多示例学习算法方面, 选择了一种基于支持向量机的多示例学习算法^[9], 记为MI-SVM; 在多任务学习算法方面, 选择了一种基于SVM的多任务学习方法^[13], 记为MtL-SVM; 在多任务多示例学习算法方面, 选择了一种基于卷积神经网络的多示例多任务分类方法^[17], 记为MtMi-CNN.

3.3 算法精度对比

分别在包含噪声的Corel和CIFAR数据集上采用MI-SVM、MtL-SVM、MtMi-CNN和MtMiAnL来训练, 本文将给出这些算法在Corel和CIFAR数据集上的精度, 如表1所示. 另外, 在Corel和CIFAR数据集中4个任务的精度结果见图1和图2.

表 1 不同数据集在各算法的总精度

Tab.1 Total accuracy of each algorithm in different data sets

	MI-SVM	MtL-SVM	MtMi-CNN	MtMiAnL
Corel	50.35	81.18	82.73	85.95
CIFAR	55.75	78.73	81.68	86.15

表1统计了各算法在Corel和CIFAR数据集上的精度,图1和图2进一步给出了在Corel和CIFAR数据集中4个任务的精度,可以看出本文MtMiAnL算法的分类效果是比较突出的。(1)相比其他3种算法,MI-SVM的分类精度较低。MI-SVM是单任务多示例学习算法,当训练样本数量较少时,其总体精度偏低。跟单任务多示例学习算法MI-SVM相比,MtMiAnL提出了多任务多示例学习算法,通过多个

相关任务的联合训练,能够在训练样本数量较少的情况下,提高各个任务的分类精度。(2)MtMiAnL获得了比MtL-SVM更高的分类精度。MtL-SVM是多任务单示例学习算法,而MtMiAnL是多任务多示例学习算法。MtMiAnL对图像进行多示例化处理后,可以降低因图像内容的多样性和歧义性所带来的分类难度,进一步提高分类精度。(3)跟MtMi-CNN相比,MtMiAnL具有更好的分类性能。虽然MtMiAnL和MtMi-CNN都是多任务多示例学习算法,但是MtMi-CNN没有考虑噪声数据的影响,而MtMiAnL通过示例加权方式,降低噪声数据对分类结果的影响,因此具有更好的分类性能。由上面各算法的结果比较得知,训练样本数量不足和噪声数据都会影响到分类预测的精度。本文算法克服了这两类问题的局限性,从而得到较好的分类精度。

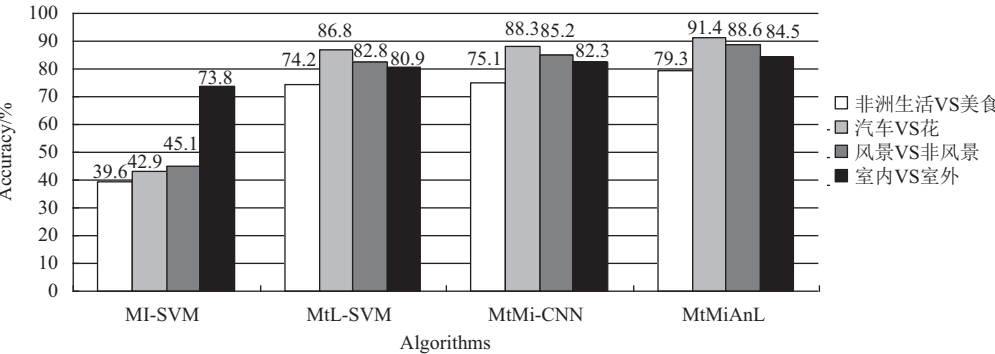


图 1 Corel数据集中不同任务的精度
Fig.1 Accuracy of each algorithm in Corel data sets

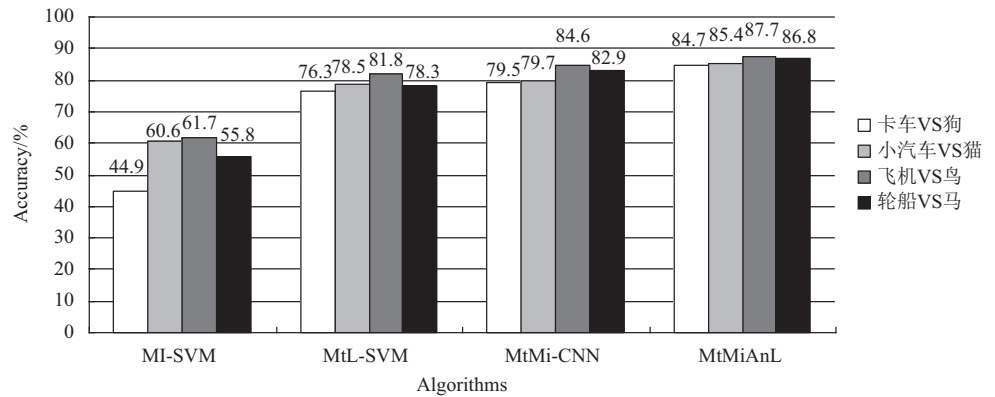


图 2 CIFAR数据集中不同任务的精度
Fig.2 Accuracy of each algorithm in CIFAR data sets

3.4 参数对精度的影响

为了研究多任务耦合参数 μ 和惩罚系数 C 对多任务分类模型的影响,本实验针对Corel和CIFAR数据集,研究在参数 μ 和 C 的变化下,MtMiAnL的分类精

度的变化。其中,多任务耦合参数 μ 的取值从[0.001, 0.01, 0.1, 0.5, 1, 5, 10, 100, 1 000]中进行选择,惩罚系数 C 的取值从 $2^{[-5:5]}$ 中进行选择。
在式(9)中,分类模型的耦合参数 μ 通过控制

w_0 和 v_i 的权值大小来调整任务间的相似性. 若令 μ 趋近无穷大时, 任务所共享的向量 w_0 将接近零向量, 此时各分类任务的相关性最低; 若令 μ 趋近于零向量时, 各分类任务所独有的向量 v_i 将接近零向量, 此时各分类任务的相关性最高. 正则化参数 C 用来控制 ζ_{ii} 使得优化问题可以容纳一定的错误分类.

图3展示了MtMiAnL算法在参数 μ 取不同值时, Corel和CIFAR数据集的分类精度; 图4展示了MtMiAnL算法在参数 C 取不同值时, Corel和CIFAR数据集的分类精度. 在实验中, 当对其中一个参数进行实验时, 另一个参数将被确定为最优值.

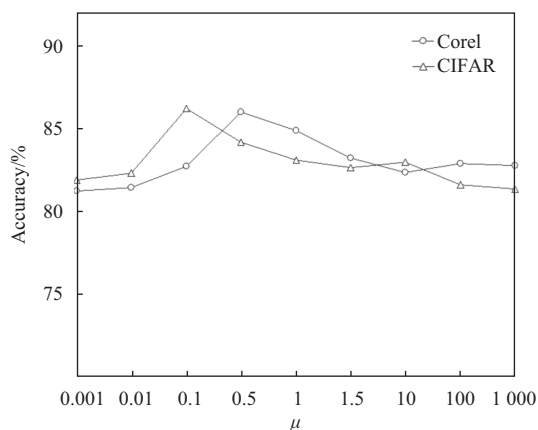


图3 Corel和CIFAR数据集在不同 μ 值下的精度

Fig.3 Accuracy of Corel and CIFAR data sets at different μ values

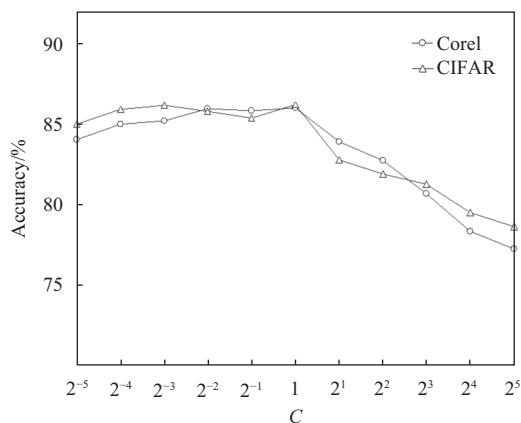


图4 Corel和CIFAR数据集在不同 C 值下的精度

Fig.4 Accuracy of Corel and CIFAR data sets at different C values

从图3可知, 在Corel数据集中, 当耦合参数 $\mu=0.1$ 时, MtMiAnL算法达到了最高的分类精度, 而当 $0.001 < \mu < 0.01$ 时, 算法的性能随着 μ 的增大而提升; 当 $\mu > 0.1$ 时, 算法的性能随着 μ 的增大而呈现递减趋势. 在CIFAR数据集中, 当耦合参数 $\mu=0.5$ 时, MtMiAnL算法达到了最高的分类精度, 当 $0.001 < \mu < 0.5$ 时, 算法的性能随着 μ 的增大提升; 而当

$\mu > 0.5$ 时, 算法的性能随着 μ 的增大而呈递减趋势. 因此, 根据此次实验, 建议 μ 的取值范围为0.1~0.5.

由图4可知, 在Corel和CIFAR数据集中, 当正则化参数 $C \leq 1$ 时, MtMiAnL算法的精度较高, 性能也比较稳定, 而当 $C > 1$ 时, MtMiAnL算法的性能下降得很快. 因此, 根据此次实验, 建议 C 的取值范围为一个较小值, 如 $C \leq 1$.

4 结束语

本文提出了一种基于抗噪声的多任务多示例学习算法, 通过对多示例包中的示例进行加权处理, 降低噪声数据在训练过程中的影响, 并对多示例加权后的目标函数进行多任务化处理, 使得所有任务的数据信息得到共享, 以提高分类精度, 从而解决了训练样本数量不充足导致分类模型效果不佳问题. 本文分别对比了3种不同的机器学习算法, 实验结果证明, 本文的方法达到更优的分类性能, 并且, 通过在实验中观察参数对分类模型的影响, 给出了参数的选取范围. 但是, 在训练过程中需要对多任务分类模型进行迭代优化, 损耗了时间. 因此, 进一步降低时间复杂度是后续研究的方向.

参考文献:

- [1] SCHÖLKOPF B, PLATT J, HOFMANN T. Multi-instance multi-label learning with application to scene classification[C]//NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge:MIT Press, 2006: 1609-1616.
- [2] 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.
- [3] CHINNIYAN K, GANGADHARAN S, SABANAİKAM K. Semantic similarity based web document classification using support vector machine [J]. International Arab Journal of Information Technology, 2017, 14(3): 258-292.
- [4] FENG S, XU D. Transductive multi-Instance multi-Label learning algorithm with application to automatic image annotation [J]. Expert Systems with Applications, 2010, 37(1): 661-670.
- [5] 梁礼欣, 郝志峰, 蔡瑞初, 等. 基于混合高斯分布伪样本生成的情感分析方法[J]. 广东工业大学学报, 2016, 33(6): 85-90.
- [6] LIANG L X, HAO Z F, CAI R C, et al. An approach to sentiment analysis of Chinese microblogs based on Gaussian mixture distribution pseudo-sample generation [J]. Journal of Guangdong University of Technology, 2016, 33(6): 85-90.
- [6] 陈炳丰, 郝志峰, 蔡瑞初, 等. 面向汽车评论的细粒度情感分析方法研究[J]. 广东工业大学学报, 2017, 34(3): 8-14.
- [6] CHEN B F, HAO Z F, CAI R C, et al. A fine-grained senti-

- [7] CHEN D Y, LIN K Y. Face-based multiple instance analysis for smart electronics billboard [J]. Multimedia Tools and Applications, 2012, 59(1): 221-240.
- [8] 杨帆, 李建平, 李鑫, 等. 基于多任务深度卷积神经网络的显著性对象检测算法[J]. 计算机应用, 2018, 38(1): 91-96.
- [9] ANDREWS S, TSOCHANTARIDIS I, HOFMANN T. Support vector machines for multiple-instance learning [J]. Advances in Neural Information Processing Systems, 2003, 15(2): 561-568.
- [10] ZHANG Q, GOLDMAN S A. EM-DD: an improved multiple-instance learning technique[C]//International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver: MIT Press, 2001: 1073-1080.
- [11] WANG J, ZUCKER J D. Solving the multiple-instance problem: a lazy learning approach[C]//Seventeenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 1119-1126.
- [12] 贾松达, 庞宇松, 阎高伟. 多任务LS-SVM在时间序列预测中的应用[J]. 计算机工程与应用, 2018, 54(3): 233-237.
- [13] JIA S D, PANG Y S, YAN G W, *et al.* Multi-task LS-SVM for application of time series prediction [J]. Computer Engineering and Applications, 2018, 54(3): 233-237.
- [14] EVGENIOU T, PONTIL M. Regularized multi-task learning[C]//Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM, 2004: 109-117.
- [15] Chih-Chung Chang, Chih-jen Lin. LIBSVM: A library for support vector machines [J]. Journal ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.
- [16] 刘成, 彭进业. 基于多任务学习的自然图像分类研究[J]. 计算机应用研究, 2012, 29(7): 2773-2775.
- [17] LIU C, PENG J Y. Research of classification method for natural images based on multitask learning [J]. Application Research of Computers, 2012, 29(7): 2773-2775.
- [18] AGGARWAL C C, YU P. A framework for clustering uncertain data streams[C]//IEEE International Conference on Data Engineering. Cancun: IEEE, 2008: 150-159.
- [19] ZENG T, JI S W. Deep convolutional neural networks for multi-instance multi-task learning [C]//IEEE International Conference on Data Mining. Atlantic: IEEE, 2015: 579-588.



· 简讯 ·

智能产业拟应用可拓创新方法扫除拦路虎,实现创新驱动发展

2018首期CAAI吴文俊人工智能科学技术奖前沿讲习班暨“领军秀”路演活动于2018年3月28日在苏州工业园区星湖街328号创意产业园A1区中国人工智能学会吴文俊人工智能科学技术奖评选基地成功举办。活动以“创新驱动，智能担当”为主题，邀请行业专家、学者与资深投资人，AI产业相关企业，行业协会，证券、银行金融嘉宾，媒体等机构代表，共同探讨人工智能的前沿发展和未来创新。本次活动由苏州工业园企业服务中心、苏州国际科技园、智博天宫（苏州）人工智能产业研究院有限公司主办，苏州广播电视台总台《创赢未来》栏目组和苏州工业园区人工智能协会协办，江苏融政科技服务有限公司承办。活动邀请了人工智能领域的专家为大家分享人工智能领域的新知识和新技术。中国人工智能学会可拓学专委会副主任兼秘书长、广东工业大学可拓学与创新方法研究所副所长李兴森教授应邀做了“可拓创新方法的应用推广”的报告，分析了智能产业发展中存在的问题，介绍了可拓创新方法的起源、方法体系及其应用情况，引起与会人员浓厚的兴趣。智博天宫（苏州）人工智能产业研究院有限公司闫海峰总经理在论坛讨论环节对可拓数据挖掘的算法应用寄予厚望。可拓创新方法及其工程应用的新技术，如可拓创意生成系统自身是一个具有良好发展前景的智能产业，同时又能为其他智能产业解决发展中的矛盾问题提供方法支撑，中国人工智能学会副秘书长、吴文俊人工智能科学技术奖办公室主任余有成团队正与工信部洽谈开展可拓创新方法及智能技术的应用培训，相关部门还诚挚邀请可拓学研究人员到苏州申报领军人才创业。期待应用可拓创新方法扫除拦路虎，实现创新驱动发展。

(广东工业大学可拓学与创新方法研究所 供稿)