

基于多示例学习的图像分类算法

汪 旗¹, 贾兆红^{1,2}

(1. 安徽大学 计算机科学与技术学院, 安徽 合肥 230601;

2. 安徽大学 计算机智能与信号处理教育部重点实验室, 安徽 合肥 230039)

摘 要:基于内容的图像分类计数通常基于图像的单一特征进行处理,而图像中包含的内容不止一个,单一的特征不足以充分描述图像,多实例学习方法由于其特殊性可以很好地解决这个难题。文中针对基于多示例学习的图像分类问题提出了一种新的多示例学习算法 DD-TSVM。该方法以图像作为包,图像中的区域作为包中示例。算法首先采用多样性密度算法寻找示例集的局部最大值以构建投影空间并将包映射为投影空间中的一个点;然后利用直推式支持向量机作为学习算法训练学习得到分类器。该算法有效地利用了未标记样本,基于 Corel 图像数据库的实验结果表明,DD-TSVM 具有良好的性能。

关键词:多示例学习;多样性密度;直推式支持向量机;图像分类

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2014)04-0088-04

doi:10.3969/j.issn.1673-629X.2014.04.022

Image Categorization by Multi-instance Learning

WANG Qi¹, JIA Zhao-hong^{1,2}

(1. College of Computer Science & Technology, Anhui University, Hefei 230601, China;

2. Ministry of Education Key Laboratory of Computing & Signal Processing, Anhui University, Hefei 230039, China)

Abstract: There have been great achievements in CBIC, and only the one single feature is generally used in the methods. Since there is more than one object in an image, it is not enough to use one feature to describe the image. The method of multi-instance learning can deal with the above problem. For the problem of multi-instance-based image categorization, a creative multi-instance learning algorithm named DD-TSVM has been proposed. This algorithm regards the image as a bag, and the region of image as an instance in the bag. First, a local maximum set has been found by diversity density algorithm to construct a projection space and transform each bag into a point in the projection space. Then, using transductive support vector machine is to get the classifier. The proposed algorithm effectively takes advantage of the unlabelled samples. The experimental results on Corel dataset show that DD-TSVM has good performance.

Key words: multi-instance learning; diverse density; transductive support vector machine; image categorization

0 引言

20世纪90年代, Dietterich^[1]等在研究药物活性预测问题时首次提出了多示例学习(Multi-Instance Learning, MIL)概念。在多示例学习问题框架中,训练集中的每个样本被看作是一个包,每个包中包含若干个示例,包具有标记而示例没有。一个包如果被标记为正,则该包中至少含有一个正示例;若某个包被标记为负,则该包中所有的示例都为负示例。经过十几年的发展,针对多示例学习问题, Dietterich等^[1]提出了三种轴平行矩形学习算法,通过在属性空间里寻找包含每个正包至少一个示例的最小矩形,预测未知包的

类别时,根据其是否有示例在矩形内判定正负。O. Maron 和 T. Lozano-Pérez^[2]提出了多样性密度(Diverse Density, DD)算法,他们认为特征空间中某点附近正包数量越多、负包示例越少,该点的多样性密度越大。算法最后求出多样性密度最大的点。Zhang 和 S. A. Goldman^[3]将期望最大算法和多样性密度算法相结合提出了 EM_DD 算法。J. Wang 和 J. D. Zucker^[4]采用 hausdorff 距离扩展 k-近邻算法提出了 Bayesian-kNN 和 Citation-kNN 算法。Ruffo^[5]修改决策树算法 C4.5 提出了 Relic 算法。S. Andrews^[6]等将支持向量机方法用于多示例学习,提出了 MI-SVM 和 mi-SVM

收稿日期:2013-03-13

修回日期:2013-07-16

网络出版时间:2014-01-28

基金项目:安徽大学青年科学研究基金项目(3305044)

作者简介:汪 旗(1985-),男,硕士研究生,研究方向为数据挖掘;贾兆红,副教授,研究方向为商务智能、数据挖掘、智能软件。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140128.1143.025.html>

算法。Chen 等^[7-8]将多示例学习问题转化为监督学习问题,他们通过构建投影空间,将包映射成空间中的点,采用支持向量机训练分类器的方法提出了两种 MIL 经典算法 DD-SVM 和 MILES。近几年,基于半监督的多示例学习受到了广泛的重视,R Rahmani 等^[9]提出了基于图的 MISSL 算法,Zhou 等^[10]提出了 missSVM 算法以及 Wang 等^[11]提出了基于图的 GMIL 算法。

图像分类需要解决三个关键的问题:

- 1) 如何表示图像;
- 2) 如何标记训练样本;
- 3) 如何进行小样本学习。

为了解决这三个问题,文中提出了一种新的直推式多示例学习算法,称之为 DD-TSVM 算法。该算法以多样性密度算法寻找到的局部极大值来构建一个新的属性空间,将包映射成属性空间中的点,最后利用直推式支持向量机 (Transductive Support Vector Machine, TSVM)^[12]技术在 MIL 框架下对图像进行分类。

1 图像分割与特征提取

给定一幅图像,采用文献[7]中提到的方法进行分割并提取图像的特征。先将图像分割成大小为 4×4 的无重叠小块,提取每个小块的颜色特征和纹理特征。其中:

(1) 颜色特征:将小块由红绿蓝的基色空间 (RGB 空间) 转换到 LUV 空间 (L 表示亮度,UV 均表示色度),取 LUV 三个分量的均值作为颜色特征,由此可得三维颜色特征向量。

(2) 纹理特征:采用 Daubechies-4 对 4×4 的小块进行一级小波变换得到 4 个频带,一个低频 LL 和三个高频 LH,HL,HH,每个频带都有 2×2 个系数。不失一般性,假设 HH 频带的系数为 $\{\alpha_{k,l}, \alpha_{k,l+1}, \alpha_{k+1,l}, \alpha_{k+1,l+1}\}$,则 HH 频带的小波特征为:

$$f = \left(\frac{1}{4} \sum_{i=0}^l \sum_{j=1}^l \alpha_{k+i,l+j}^2 \right)^{\frac{1}{2}} \quad (1)$$

LH,HL 的小波特征可参照 HH 求出,由此可以得到 3 维纹理特征向量。

采用 k-means 方法将分割后的图像小块聚成若干类,每一类对应着图像的一个区域,则区域 R_i 可表示为 f_i 。

(3) 形状特征:对图像的任意一个区域 R_i ,运用阶数为 1,2,3 的归一化惯性因子来描述其形状特征。惯性因子 γ 的公式如下:

$$I(R_i, \gamma) = \frac{\sum_{r \in R_i} \|r - \hat{r}\|^\gamma}{V_i^{1+\frac{\gamma}{2}}} \quad (2)$$

其中, \hat{r} 为区域 R_i 的质心; r 为区域 R_i 中每个像素的坐标; V 为区域 R_i 的像素总量。

在二维空间里,分割区域的最小外接圆的惯性因子最小,设最小外接圆的第 γ 个惯性因子为 I_γ ,区域 R_i 的三维形状特征可由下式计算得到

$$s_i = \left[\frac{I(R_i, 1)}{I_1}, \frac{I(R_i, 2)}{I_2}, \frac{I(R_i, 3)}{I_3} \right]^T \quad (3)$$

这样,一幅被分成 n 个区域的图像 B ,可被表示为 $B = \{x_i; i = 1, 2, \dots, n\}$, $x_i = [\hat{f}_i^T, s_i^T]^T$ 对应着图像区域 R_i 的 9 维的特征向量。

2 DD-TSVM 算法

2.1 构造投影空间

已标记样本集 $L = \{(B_1, y_1), (B_2, y_2), \dots, (B_{|L|}, y_{|L|})\}$, $y_i \in \{-1, +1\}$, $i = 1, 2, \dots, |L|$, $+1$ 表示属于本类的样本, -1 表示不属于本类的样本; 未标记样本集 $U = \{B_1, B_2, \dots, B_{|U|}\}$ 。DD 算法主要基于以下设想:在属性空间中某点附近的正包越多,反包示例越少,该点的多样性密度越大。在已标记的样本集中,令 B_i^+ 代表第 i 个正包, B_{ij}^+ 为第 i 个正包的第 j 个示例, B_{ijk}^+ 为第 i 个正包的第 j 个示例的第 k 个属性的值;令 B_i^- 代表第 i 个负包, B_{ij}^- 为它的第 j 个示例, B_{ijk}^- 为 B_{ij}^- 的第 k 个属性的值;令 d 代表属性空间中多样性密度最大的点,寻找该点的目标函数可以定义为:

$$\arg \max_x \Pr(x = d | B_1^+, \dots, B_n^+, B_1^-, \dots, B_m^-) \quad (4)$$

假定包与包之间是独立的,根据 Bayes 理论,多样性密度可由下式来确定:

$$d = \arg \max_x \prod_i \Pr(x = d | B_i^+) \prod_i \Pr(x = d | B_i^-) \quad (5)$$

在实际使用上述多样性密度时,还需要对公式 (5) 中的乘积部分进行例化,如 O. Maron 等人采用了 noisy-or 模型进行例化:

$$\Pr(x = d | B_i^+) = 1 - \prod_j (1 - \Pr(x = d | X_{ij}^+)) \quad (6)$$

$$\Pr(x = d | B_i^-) = \prod_j (1 - \Pr(x = d | X_{ij}^-)) \quad (7)$$

他们将示例出现在目标位置的概率定义为该示例与潜在目标之间的距离函数,即:

$$\Pr(x = t | B_{ij}) = \exp(-\|B_{ij} - x\|^2) \quad (8)$$

因为存在多个局部极小值,因此在使用梯度下降法来寻找目标点时要将每一个正包示例都作为初始点进行第一次搜索。

2.2 包特征匹配

假设有投影特征空间 $\Theta = \{x_1', x_2', \dots, x_k'\}$, 其中 x_i'

为投影空间的第 i 维。包 $B_i = \{x_{ij} \mid j = 1, 2, \dots, n_i\}$ 到投影空间中第 i 维的距离可由下式来计算:

$$s(x'_i, B_i) = \max_{j=1,2,\dots,n_i} \exp(-\|x_{ij} - x'_i\|^2) \quad (9)$$

包在投影空间中的映射结果定义如下:

$$\varphi(B_i) = [s(x'_1, B_i), s(x'_2, B_i), \dots, s(x'_k, B_i)] \quad (10)$$

若特征空间为 N 维, 匹配后的包可视为 N 维特征空间里的一个点。

2.3 TSVM

由于 SVM 只能用已经标记的样本来训练分类器, 为了既能利用无标记样本又能保证最后的泛化能力, Joachims^[12]在 SVM 的基础上引入对无标记样本的惩罚因子, 提出了 TSVM 并将之用于文本分类。下面简要介绍 TSVM 的原理:

在已标记样本集 $\{(x_i, y_i) \mid i = 1, 2, \dots, l, y_i \in \{+1, -1\}\}$ 和无标记的样本集 $\{x_j^* \mid j = 1, 2, \dots, k\}$ 上最大化 margin, 其目标函数如下:

$$\min_{w, b, \xi, \xi^*, y_j^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i + D \sum_{j=1}^k \xi_j^* \quad (11)$$

$$\text{s. t. } y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, 2, \dots, l$$

$$y_j^*(w^T x_j^* + b) \geq 1 - \xi_j^*$$

$$\xi_j^* \geq 0, j = 1, 2, \dots, k$$

$$y_j^* \in \{+1, -1\}$$

其中, $\xi = (\xi_1, \xi_2, \dots, \xi_l, \xi_1^*, \xi_2^*, \dots, \xi_k^*)^T$ 为松弛变量; C 和 D 分别为已标记样本的惩罚因子和无标记样本的惩罚因子。

TSVM 最终会计算出一个分类超平面 w^* 和无标记样本的类别标号 $y_j^* \in \{-1, +1\}, j = 1, 2, \dots, k$ 。假设求出的最优解为 w^* , 包分类器可以定义如下:

$$\text{lable}(B) = \text{sign}(w^{*T} \varphi(B)) \quad (12)$$

经过上述分析, DD-TSVM 算法步骤可描述如下:

输入: 有标记的图像集合 L 、没有标记的图像集合 U ;

输出: TSVM 分类器 w^* 和没有标记图像的标记 $\{y_i^*\}_{i=1}^{|U|}$ 。

Step1 生成包和示例: $\forall B \in L \cup U$, 采用分割图像, 提取底层特征。假设包 B_i 被分成 n_i 个区域, 则包 $B_i = \{x_{ij} \mid j = 1, 2, \dots, n_i\}$, 其中 x_{ij} 表示包 B_i 的第 j 个示例。

Step2 构造投影空间 Θ : 将所有有标记包中的示例集合在一起, 组成示例集 $I = \{x_l \mid l = 1, 2, \dots, n\}$, $n = \sum_{i=1}^{|L|} n_i$, 其中 n 为示例总量, n_i 为第 i 个包中示例数量;

采用 DD 算法, 输出全部局部极值 x'_i , 利用局部极值来构造投影空间 $\Theta = \{x'_1, x'_2, \dots, x'_k\}$ 。

Step3 计算投影特征: 初始化 $R = \emptyset; \forall B_i \in L$, 计算包 B_i 的投影特征 $\varphi(B_i)$ 并将 $(\varphi(B_i), y_i)$ 加入 R (y_i 为对应包 B_i 的标签); $\forall B_j \in U$, 计算包 B_j 的投影特征 $\varphi(B_j)$ 并将 $(\varphi(B_j), 0)$ 加入 R (0 表示该样本无标记)。

Step4 训练分类器: 在 R 上求公式 (11) 的最优解, 返回 TSVM 分类器 w^* 和没有标记图像的标签 $\{y_i^*\}_{i=1}^{|U|}$ 。

3 实验结果与分析

实验采用 Corel-1k 图像数据集, 共 10 组, 分别为: African people and villages, Beach, Historical Building, Buses, Dinosaurs, Elephants, Flowers, Horses, Moutains and glaciers, Food。每一组包含 100 幅格式为 JPEG 的彩色图片, 图片分辨率为 384×256 或者 256×384 。采用第 2 节阐述的方法对图像进行处理, 提取底层特征生成包和示例。

由于 Corel 数据集的图像类别多于两个, 所以实验采用 one-against-the-rest 方法, 即对每一类图像训练一个区分该类与其他类图像的分类器。实验平台为: Core I3 处理器, 2 G 内存, Win7 操作系统以及 Matlab2010 仿真环境。

为了验证 DD-TSVM 算法的有效性, 将其与经典的 DD-SVM 和 MILES 算法进行对比实验。在 DD-TSVM 算法实验中, 对每一类别的图像, 随机选取 50 幅图像加入训练集, 剩下的 50 幅图像作为测试样本加入测试集, 训练集中的图像 20 幅标记为正, 30 幅不标记; 在其他类别图像中随机选取 100 幅图像, 50 幅加入训练集, 剩下的 50 幅图像加入测试集。加入训练集的其他类图像中 20 幅标记为负, 30 幅不标记。每个实验重复十次, 取平均精度值, 最终结果如表 1 所示。

由表 1 可见, 文中提出的 DD-TSVM 算法在总的平均分类精度上明显高于 DD-SVM 算法和 MILES 算法, 10 类图像中有 7 类的分类精度都是最高的, 只有 3 类图像的分类精度稍低, 这是因为 TSVM 能够利用训练集中未标记样本包含的信息来训练分类器, 有效地解决了训练样本不足的问题, 并且 TSVM 比 MILES 算法中使用的 1-norm SVM 具有更好的学习分类能力。

4 结束语

文中利用多样性密度 (DD) 和直推式支持向量机 (TSVM) 算法, 提出了一种新的多示例学习算法 (DD-TSVM 算法)。该算法采用 DD 算法寻找样本集中的

表 1 基于 Corel 图像集的分类精度值 %

图像类别	DD-SVM	MILES	文中方法
Africa	66.70	68.8	78.0
Beach	68.4	66.0	81.0
Buildings	74.3	75.7	80.9
Buses	90.3	90.3	93.8
Dinosaurs	99.7	99.7	92.4
Elephants	76.0	77.7	80.1
Flowers	88.3	96.4	92.3
Horses	94.3	95.0	93.6
Mountains	70.3	71.0	78.0
Food	87.0	85.4	88.2
平均值	81.63	82.6	85.83

注:加粗的为在各类上得到的最好分类精度。

局部极大值来构建特征空间,然后采用 TSVM 训练分类器。文中算法特点在于运用 DD 算法能够找出样本集中最特殊的示例,运用直推式支持向量机能够有效利用未标记的样本来训练分类器,在基于 Corel 图像数据库的实验结果表明,文中算法是一种有效的图像分类算法。

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] Maron O, Lozano-Pérez T. A framework for multiple-instance learning[C]//Proceedings of the 1997 conference on advances in neural information processing systems 10. Cambridge, MA, SA: MIT Press, 1998: 570-576.

[3] Zhang Q, Goldman S A. EM-DD: An improved multiple-instance learning technique[C]//Proc of advances in neural information processing systems 14. Cambridge, MA, USA: MIT

Press, 2002: 1073-1080.

[4] Wang J, Zucker J D. Solving the multiple-instance problem: A lazy learning approach[C]//Proceedings of the 17th international conference on machine learning. San Francisco, CA: [s. n.], 2000: 1119-1125.

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

[6] Andrews S, Hofmann T, Tsochantaridis I. Multiple instance learning with generalized support vector machines[C]//Proceedings of the 18th national conference on artificial intelligence. Edmonton, Canada: AAAI Press, 2002: 943-944.

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

[8] Chen Yixin, Bi Jinbo, Wang J Z. MILES: Multiple-instance learning via embedded instance selection[J]. IEEE trans on pattern analysis and machine intelligence, 2006, 28(12): 1931-1947.

[9] Rahmani R, Goldman S A. MISSL: Multiple instance semi-supervised learning[C]//Proc of the 23rd int conf on machine learning. Pittsburgh: [s. n.], 2006: 705-712.

[10] Zhou Z H, Xu J M. On the relation between multi-instance learning and semi-supervised learning[C]//Proc of the 24th ICML. Corvalis: [s. n.], 2007: 1167-1174.

[11] Wang C H, Zhang L, Zhang H J. Graph-based multiple-instance learning for object-based image retrieval[C]//Proc of the 1st ACM int conf on multimedia information retrieval. Vancouver: [s. n.], 2008: 156-163.

[12] Joachims T. Transductive inference for text classification using support vector machines[C]//Proc of the 16th international conference on machine learning. San Francisco, USA: [s. n.], 1999: 200-209.

(上接第 87 页)

n.], 2011.

[2] 王镇道, 陈迪平, 文康益. JPEG 图像压缩算法的 IP 核设计[J]. 计算机应用, 2005, 25(5): 1076-1077.

[3] 潘 榕, 刘 昱, 侯正信, 等. 基于局部 DCT 系数的图像压缩感知编码与重构[J]. 自动化学报, 2011(6): 674-681.

[4] 万 源, 吴传生. MPEG-4 视频 DCT 量化模块的改进方法[J]. 计算机工程与应用, 2007, 43(12): 42-44.

[5] 安会乐. JPEG/MJPEG 中 Huffman 编解码的 IP 设计[D]. 西安: 西安电子科技大学, 2007.

[6] 李 薇, 胡伟文, 沈 静. 霍夫曼编码和游程编码在图像编码中的应用[J]. 舰船电子工程, 2010, 30(7): 67-69.

[7] 沈 洁, 杜宇人, 殷玲玲, 等. 基于 DCT 变换的图像压缩技术研究[J]. 信息技术, 2006(10): 133-134.

[8] 武 瑛. DCT 变换在图像压缩中的应用[J]. 计算机与现代化, 2013(4): 103-106.

[9] 张 羽, 梁德群. 基于图像分割改进的 DCT 压缩算法[C]//第十二届全国图象图形学学术会议论文集. 出版地不详: 中国图象图形学学会, 2005.

[10] Zhao Ming, Liu Zhibin, Ren Baosheng, et al. Improved ANN algorithm based on the change of search direction[C]//Proceedings of the 2010 international conference on computational and information sciences. [s. l.]: [s. n.], 2010.

[11] 涂望明, 魏友国, 施少敏. MATLAB 在数字图像处理中的应用[J]. 微计算机信息, 2007, 23(6): 299-300.

[12] 沙 亮, 王贵锦, 赵安邦, 等. H. 264 编码下基于 DCT 系数量化值的 PSNR 估计方法[J]. 清华大学学报(自然科学版), 2008(4): 510-513.

基于多示例学习的图像分类算法

作者:	汪旗, 贾兆红, WANG Qi, JIA Zhao-hong
作者单位:	汪旗, WANG Qi(安徽大学 计算机科学与技术学院, 安徽 合肥, 230601), 贾兆红, JIA Zhao-hong(安徽大学 计算机科学与技术学院, 安徽 合肥 230601; 安徽大学 计算机智能与信号处理教育部重点实验室, 安徽 合肥 230039)
刊名:	计算机技术与发展
	<div>ISTIC</div>
英文刊名:	Computer Technology and Development
年, 卷(期):	2014(4)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjfz201404022.aspx