

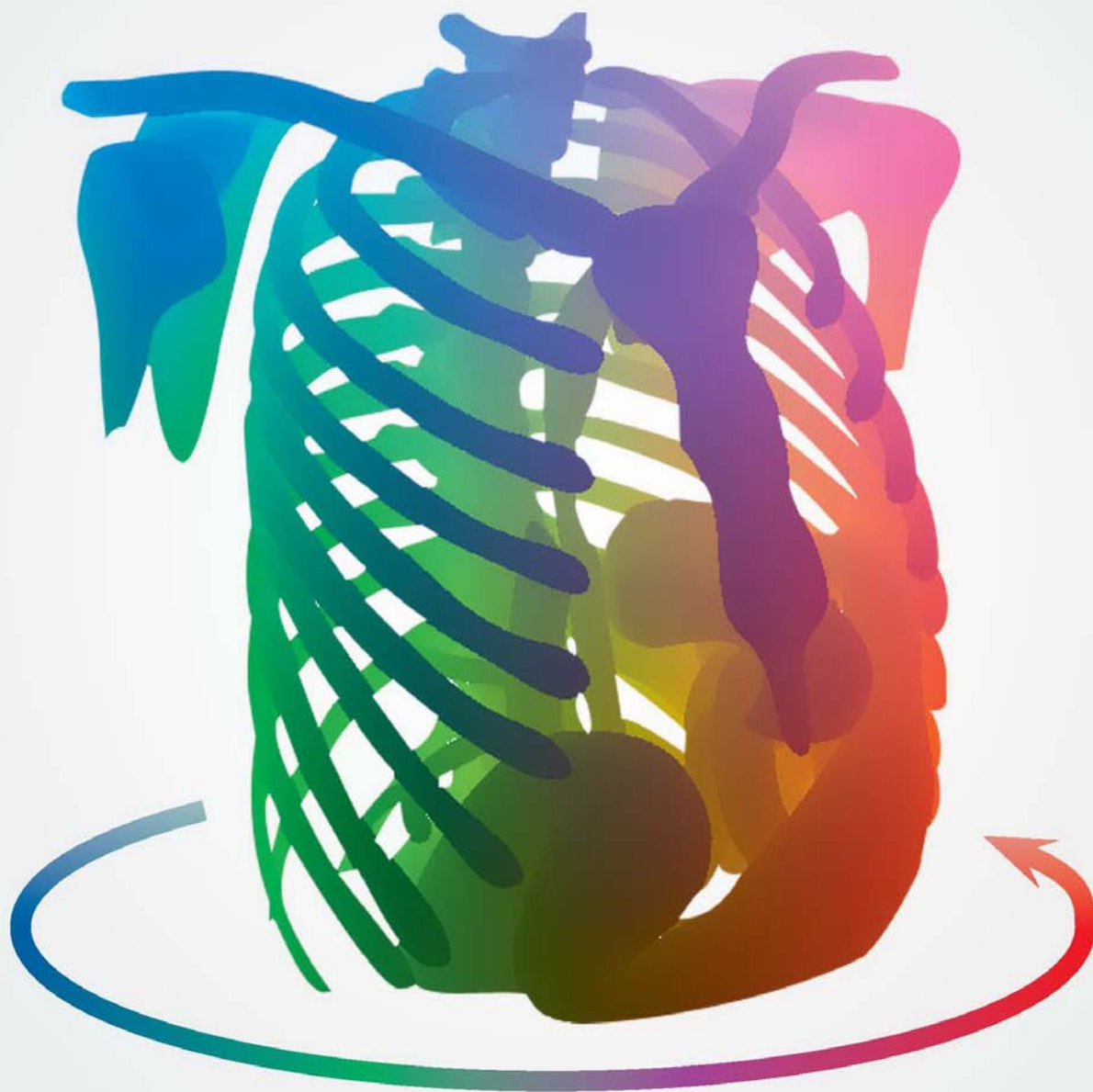
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
02
VOL.20

ISSN1006-8961
CN11-3758/TB



等值面光线跟踪 P193



第20卷第2期（总第226期）

2015年2月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院遥感与数字地球研究所

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 李小文

编辑出版《中国图象图形学报》编辑出版委员会

邮政信箱 北京9718信箱

邮 编 100101

电子信箱 jig@irsa.ac.cn

电 话 010-64807995

网 址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

（邮政信箱：北京399信箱 邮编：100048）

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Institute of Remote Sensing and Digital Earth, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen

Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics

P.O.Box 9718, Beijing, P.R.China

Zip code 100101

E-mail jig@radi.ac.cn

Telephone 010-64807995

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China))

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

ISSN 1006-8961

CODEN ZTTXFZ

国外发行代号 M1406

国内邮发代号 82-831

国内定价 50.00元



多尺度活动网格在云场景仿真中的应用(第0202页)



自然环境视频中基于显著鲁棒轨迹的行为识别(第0245页)



虚拟手抓持力觉生成算法真实性的评价(第0280页)

图像处理和编码

引入神经网络中间神经元的快速小波图像压缩

张海涛, 张永霖 0159

拟正态分布扩散的图像平滑

周先春, 汪美玲, 周林锋 0169

可选特征的快速分形图像编码

袁宗文, 鲁业频, 杨汉生 0177

图像分析和识别

融合多尺度码本的全局编码图像分类

董振宇, 赵杰煜, 祝军 0183

计算机图形学

结合K-D树和Shell的快速动态等值面光线跟踪法

罗月童, 石放放, 张伟, 朱会国 0193

多尺度活动网格在云场景仿真中的应用

范晓磊, 张立民, 张建廷, 徐涛 0202

第十届和谐人机环境联合会议专栏

低资源语言的无监督语音关键词检测技术综述

杨鹏, 谢磊, 张艳宁 0211

快速离散Curvelet变换域的图像融合

杨勇, 董松, 黄淑英 0219

高效视频编码中变换跳过模式的快速选择

王宁, 张永飞, 樊锐 0229

主动学习的多标签图像在线分类

徐美香, 孙福明, 李豪杰 0237

自然环境视频中基于显著鲁棒轨迹的行为识别

易云, 王瀚漓 0245

特征变换和数据集分块的行人比对

沈洋, 林巍晓, 殷惠清 0254

移动手持环境下触摸式目标选择技术对比

辛义忠, 李岩, 袁伟强, 郑谦 0264

虚拟环境的用户意图捕获

程成, 赵东坡, 卢保安 0271

虚拟手抓持力觉生成算法真实性的评价

杨文珍, 许艳, 颜传武, 吴新丽, 张昊, 孟闯 0280

交互式乐器演奏的六自由度力觉渲染方法

童号, 史有姣, 王党校, 张玉茹 0288

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Multi-scale moving grids and its application in clouds scene simulation(P0202)



Action recognition from unconstrained videos via salient and robust trajectory (P0245)



Validation the Authenticity of Virtual fingers Grasping Haptic Algorithm(P0280)

Image Processing and Coding

- Fast wavelet image compression introducing neural network of relay neuron
Zhang Haitao, Zhang Yonglin 0159
- Image smoothing algorithm based on matching normal distribution diffusion
Zhou Xianchun, Wang Meiling, Zhou Linfeng 0169
- Optional features fast fractal image coding algorithm
Yuan Zongwen, Lu Yepin, Yang Hansheng 0177

Image Analysis and Recognition

- Image classification based on global coding combined with multi-scale codebook
Dong Zhenyu, Zhao Jieyu, Zhu Jun 0183

Computer Graphics

- Fast isosurface ray tracing method by combining K-D tree and Shell
Luo Yuetong, Shi Fangfang, Zhang Wei, Zhu Huiguo 0193
- Multi-scale moving grids and its application in clouds scene simulation
Fan Xiaolei, Zhang Limin, Zhang Jianting, Xu Tao 0202

Column of HHME'2014

- Survey on unsupervised spoken term detection for low-resource languages
Yang Peng, Xie Lei, Zhang Yanning 0211
- Image fusion based on fast discrete Curvelet transform
Yang Yong, Tong Song, Huang Shuying 0219
- Fast transform skip mode decision for high efficiency video coding
Wang Ning, Zhang Yongfei, Fan Rui 0229
- Online multi-label image classification with active learning
Xu Meixiang, Sun Fuming, Li haojie 0237
- Action recognition from unconstrained videos via salient and robust trajectory
Yi Yun, Wang Hanli 0245
- Pedestrian re-identification based on feature transformation and dataset classification
Shen Yang, Lin Weiyao, Yin Huiqing 0254
- Comparison of target selection techniques in mobile handheld environment
Xin Yizhong, Li Yan, Yuan Weiqiang, Zheng Qian 0264
- Intent understanding for virtual environments
Cheng Cheng, Zhao Dongpo, Lu Baoan 0271
- Validation the authenticity of virtual fingers grasping haptic algorithm
Yang Wenzhen, Xu Yan, Yan Chuanwu, Wu Xinli, Zhang Hao, Meng Chuang 0280
- Interactive simulation of music instrument playing by a six degree-of-freedom haptic rendering approach
Tong Hao, Shi Youjiao, Wang Dangxiao, Zhang Yuru 0288

中图法分类号: TP37 文献标识码: A 文章编号: 1006-8961(2015)02-0237-08

论文引用格式: Xu M X, Sun F M, Li H J. Online multi-label image classification with active learning[J]. Journal of Image and Graphics, 2015, 20(2): 0237-0244. [徐美香, 孙福明, 李豪杰. 主动学习的多标签图像在线分类[J]. 中国图象图形学报, 2015, 20(2): 0237-0244.] [DOI:10.11834/jig.20150210]

主动学习的多标签图像在线分类

徐美香¹, 孙福明¹, 李豪杰²

1. 辽宁工业大学电子与信息工程学院, 锦州 121001; 2. 大连理工大学软件学院, 大连 116300

摘要: 目的 在多标签有监督学习框架中, 构建具有较强泛化性能的分类器需要大量已标注训练样本, 而实际应用中已标注样本少且获取代价十分昂贵。针对多标签图像分类中已标注样本数量不足和分类器再学习效率低的问题, 提出一种结合主动学习的多标签图像在线分类算法。方法 基于 min-max 理论, 采用查询最具代表性和最具信息量的样本挑选策略主动地选择待标注样本, 且基于 KKT (Karush-Kuhn-Tucker) 条件在线地更新多标签图像分类器。结果 在 4 个公开的数据集上, 采用 4 种多标签分类评价指标对本文算法进行评估。实验结果表明, 本文采用的样本挑选方法比随机挑选样本方法和基于间隔的采样方法均占据明显优势; 当分类器达到相同或相近的分类准确度时, 利用本文的样本挑选策略选择的待标注样本数目要明显少于采用随机挑选样本方法和基于间隔的采样方法所需查询的样本数。结论 本文算法一方面可以减少获取已标注样本所需的人工标注代价; 另一方面也避免了传统的分类器重新训练时利用所有数据所产生的学习效率低下的问题, 达到了当新数据到来时可实时更新分类器的目的。

关键词: 多标签分类; 主动学习; 在线学习; min-max 理论

Online multi-label image classification with active learning

Xu Meixiang¹, Sun Fuming¹, Li haojie²

1. School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China;

2. School of Software, Dalian University of Technology, Dalian 116300, China

Abstract: **Objective** Supervised machine learning methods have been applied to multi-label image classification problems with tremendous success. Despite the different learning mechanisms, the performances of these methods heavily rely on the number of labeled training examples. However, the acquisition of labeled examples requires significant efforts from annotators, which hinders the application of supervised learning methods to large-scale problems. In this paper, we extend an active querying method driven by informativeness and representativeness in single-label learning to multi-label image classification. Given that the training set grows in active learning, the classifier needs to be updated accordingly. A new classifier is required to use all training samples for re-training. Under such condition, the training burden of the classifier increases significantly. A highly effective online learning algorithm is explored to speed up learning efficiency. To deal with the massive multi-label classification problems, a novel algorithm named active learning with informativeness and representativeness for online multi-label learning (AIR-OML), which aims at sample selection strategy and at update issues in classifications, is presented. **Method** The sample selection strategy in active learning is based on the min-max theory, querying the most

收稿日期: 2014-07-25; 修回日期: 2014-09-25

基金项目: 国家自然科学基金项目 (61272214, 61173104)

第一作者简介: 徐美香 (1989—), 女, 辽宁工业大学通信与信息系统专业在读硕士研究生, 主要研究方向为图像语义理解。

E-mail: xumx0721@sina.com

Supported by: National Natural Science Foundation of China (61272214, 61173104)

informative and the most representative examples to retrain the multi-label learner. Kullback-Leibler divergence and Karush-Kuhn-Tucker conditions are utilized to update the multi-label classifier online in real time. Combining active learning with online learning, we propose the AIR-OML algorithm. **Result** Experiments are conducted in four different real-world datasets with four evaluation criteria to evaluate the presented algorithm. Experimental results demonstrate that the sample selection strategy explored has a significant advantage over the other two existing sample selection strategies, and the classifier can achieve the best performance with fewer new samples by querying the most informative and representative examples. **Conclusion** The AIR-OML algorithm can reduce the cost of human annotation and realize the goal of updating the classifier timely when newly labeled examples arrive.

Key words: multi-label classification; active learning; online learning; min-max theory

0 引言

分类问题是模式识别的核心研究内容,其目的是利用已标注的样本,通过学习或训练设计出一个分类器,然后利用这个分类器来预测新样本的标签。在传统的分类问题中,一个对象仅用一个示例表示,且该示例只对应一个类别,即单标签学习。但是,在现实世界的许多应用中,一个样本往往和多个类别相关联,例如在场景分类中,一幅图像可以同时拥有多个概念,如“树”、“大象”、“狮子”和“草地”等。因此,解决此类多义性问题,需要引入多标签学习框架^[1]。

在多标签学习框架下,从监督的角度上来看,算法可以划分为有监督方法和半监督方法。在有监督学习方法中,分类器需要使用大量已标注样本进行学习,以确保最终训练出的学习器具有较强的泛化性能。然而,获取大规模的训练样本集,利用人工标注不仅费时费力且代价相当昂贵。同时,训练样本过多也会大大降低分类器的训练效率。因为相比于已标注样本来说,未标注样本数量巨大且容易获取。因此,如何综合利用少量已标注样本和大量未标注样本来提高分类器的学习性能是当前机器学习和模式识别的重要研究内容之一。

利用未标注样本进行机器学习的算法又可归纳为3类:半监督学习、直推式学习和主动学习,这些算法在单标签学习问题中已经得到了广泛的研究和应用。其中,主动学习由于能够自主地选择最有价值的样本交由“神谕”进行标注,大大减少了人工标注的代价,受到了越来越多研究者的关注,并且在单标签学习中已经得到了广泛应用。近年来,学者们开始将它们引入到多标签学习的框架中开展研究。

然而,由于多标签学习本身的复杂性,使得主动学习应用在多标签学习框架中具有一定的局限性。

本文的研究内容之一是研究主动学习在多标签学习框架中应用问题。根据文献[2]中的 min-max 理论,Huang 等人^[3]提出了一种基于查询最具信息量和最具代表性的样本的主动学习方法用于解决单标签学习问题。本文将此种方法推广至多标签学习框架中。另外,使用主动学习挑选样本,使训练样本集合的规模不断地扩大,导致分类器也需要及时更新。通常,传统的分类器进行重训练时,需要使用全部的训练样本,这样大大增加了大量训练数据的存储负担。而当训练数据量较大时,分类器的训练时间耗时相当巨大。针对这一问题,本文的另一研究内容是分类器的在线学习问题。即研究如何仅利用新标注的样本对已有的分类器进行实时更新的方法,力图克服存储大量训练样本带来的空间问题,也克服传统的分类器再学习利用全部样本产生的学习效率低下的问题,达到新样本到来时可以实时更新分类器的目的。

本文的主要贡献如下:

- 1) 将一种基于查询最具信息量和最具代表性样本策略的单标签主动学习方法推广应用于多标签学习;
- 2) 采用多标签在线学习算法,仅利用新数据实时更新分类器,大大提高了分类器的再学习效率,显著降低了训练数据的存储要求。

1 相关工作

主动学习是同时利用已标注数据和未标注数据的信息来提升学习器性能的主流技术之一。其目的是在保证分类器性能的前提下,尽可能使用

较少的训练样本,最大限度地降低标注代价。通常情况下,主动学习由学习引擎和选择引擎两个阶段组成。其中,学习引擎负责在已给的标注样本集上利用已有的监督学习方法构造基准分类器。而选择引擎则负责在未标注的样本集中挑选样本,经“神谕”标注后,将挑选的样本及其标签交给学习引擎进行学习。

根据未标注样本获取手段的不同,可以把主动学习大致划分为3种模型:基于成员查询的主动学习、基于流的主动学习和基于池的主动学习。本文所研究的方法归属于基于池的方法。在基于池的主动学习中,有两种最典型的样本选择策略:基于信息量的选择策略和基于代表性的选择策略。基于信息量的选择策略从大量未标注样本中查询信息量最丰富的样本。其典型方法包括不确定采样^[4]、基于委员会查询^[5]和最优实验设计^[6]。但这些方法存在着一个共同缺点是不能充分利用大量的未标注样本,并且仅由少数的已标注样本来决定选择哪些样本,致使得到的样本集往往带有偏差。基于代表性的选择策略通常采用聚类的方法,它主要是利用未标注样本的聚类结构。其主要缺点是,算法的性能在很大程度上取决于聚类结果的好坏。另外,还有一些样本选择方法是将上述两种策略结合起来,即将查询最具代表性样本的方法和查询最具信息量样本的方法结合起来寻找最优待标注样本^[7-9]。特别地,文献[7]提出了一种采样方法,该方法同时利用了聚类信息和未标注样本的分类间隔。为了更好地查询选择样本,Donmez等人^[8]推广了文献[9]中的主动学习方法,它将样本的不确定性和密度有机地结合起来,取得了一定的成效。

近年来,主动学习在单标签学习问题得到了广泛的应用,并且在多标签学习问题中也开始得到人们的普遍关注。有学者将其直接或融合其他方法(如半监督方法)应用到多标签分类问题中,得到了较好的学习性能。Li等人^[10]将支持向量机和主动学习相结合应用于多标签图像分类,取得了比单纯利用支持向量机处理多标签分类问题更好的效果。齐国君^[11-12]在关联多标签学习方法中提出挑选样本一概念对用于人工标注,该方法在考虑了待选样本一概念对的不确定性的同时,也考虑了概念的准确性。蒋华等人^[13]将球结构支持向量机与主动学习方法结合用于多标签分类,依据球重叠区域样本

距离差值确定样本类别,分析多标签分类特性,采用样本近邻方法更新分类器,用较少的训练样本获得更有效的分类结果。Zhang等人^[14]结合主动学习机制和多视图学习,提出了一种多视图多标签主动学习算法。该算法首先在每个视图上分析了样本和标签的不确定性,然后基于多视图融合理论获取不同视图的不确定性,最后根据在样本、标签和视图层面上整体的不确定性来选择最具信息量的样本标签对。受版本空间策略的启发,文献[15]将基于池的主动学习算法推广至基于“一对多”二分类分解框架的多标签分类问题中使用,取得了一定的成效。刘瑞阳等人^[16]提出基于平均期望间隔的多标签分类主动学习方法,通过计算支持向量机分类器中的期望间隔作为样本选择标准。许钰提出一种基于偏侧分类间隔的SVM主动学习选择策略^[17],选择位于分类超平面一侧分类间隔中间的正类样本进行标记,加入训练集训练分类器,取得了较好的效果。

本文主要研究主动学习在多标签学习框架下的应用和多标签分类器在线更新问题。首先介绍基于min-max理论的最大间隔学习。然后将基于查询最具信息量和最具代表性样本的主动学习方法QUIRE^[3](query informative and representative example)从单标签学习推广至多标签学习框架中。在此基础上,结合在线学习^[12],提出一种结合主动学习的多标签在线学习算法AIR_OML(active learning with informativeness and representativeness for online multi-label learning)算法。最终,在多标签数据集上的实验结果验证了该算法的有效性。

2 理论及算法

2.1 理论

令 $D = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_{n_l}, \mathbf{y}_{n_l}), \mathbf{x}_{n_l+1}, \dots, \mathbf{x}_n\}$ 表示整个数据集,包括已标注数据和未标注数据。假设数据集中前 n_l 个样本是已标注的,剩下 $n_u = n - n_l$ 个样本是未标注的。每个样本由一个 d 维向量 $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$ 来表示,其标签向量为 $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{ic}]^T$, c 为类别总数。每次从池中选择一个样本 $\mathbf{x}_s$ 并查询其标签。为简单起见,把数据集 D 划分为3个部分:已标注样本集 D_l 、当前选择的样本 $\mathbf{x}_s$ 和未标注样本集 D_u 。

2.1.1 基于 min-max 理论的最大间隔主动学习

令 h^* 表示由已标注样本训练得到的分类模型,即

$$h^* = \arg \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \sum_{i=1}^{n_l} l(y_i, h(x_i)) \quad (1)$$

式中, H 表示核函数 $\kappa(\cdot, \cdot)$ 的再生希尔伯特空间: $\mathbf{R}^d \times \mathbf{R}^d \rightarrow \mathbf{R}$, $l(z)$ 为损失函数。给定分类器 h^* , 基于最大间隔的方法是选择距离决策界最近的未标注样本,即

$$s^* = \arg \min_{n_l < s \leq n} |h^*(x_s)| \quad (2)$$

根据文献[18],式(2)可近似为

$$s^* = \arg \min_{n_l < s \leq n} f(D_l, x_s) \quad (3)$$

式中, $f(D_l, x_s)$ 定义为

$$f(D_l, x_s) = \max_{y_s = \pm 1} \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \sum_{i=1}^{n_l} l(y_i, h(x_i)) + l(y_s, h(x_s)) \quad (4)$$

式(3)也可以写成下面的最小最大形式

$$\min_{n_l < s < n, y_s = \pm 1} \max E(D_l, x_s)$$

式中, $E(D_l, x_s) = \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \sum_{i=1}^{n_l} l(y_i, h(x_i)) + l(y_s, h(x_s))$ 。

根据主动学习的 min-max 理论,当前选择的样本 x_s 可以确保目标函数具有较小的值。为了选择同时满足最具信息量和最具代表性的样本,将目标函数 $f(D_l, x_s)$ 进行扩展,使其包含所有的未标注样本。若已知 D_u 中未被选择的未标注样本的类别标签 y_u ,则评价函数 $f(D_l, x_s)$ 可以改写为

$$f(D_l, D_u, y_u, x_s) = \max_{y_s = \pm 1} \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \sum_{i=1}^n l(y_i, h(x_i)) \quad (5)$$

式中, y_u 是未知的。要想求得 y_u 的最优解,根据流形假设, $f(D_l, D_u, y_u, x_s)$ 应取得一个较小的值。通过最小化 $f(D_l, D_u, y_u, x_s)$,可以得到 y_u 的近似解,进而得到选择样本的评价函数

$$\begin{aligned} \hat{f}(D_l, D_u, x_s) &= \min_{y_u \in \{\pm 1\}^{n_u-1}} f(D_l, D_u, y_u, x_s) = \\ &= \min_{y_u \in \{\pm 1\}^{n_u-1}} \max_{y_s = \pm 1} \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \sum_{i=1}^n l(y_i, h(x_i)) \end{aligned} \quad (6)$$

2.1.2 求解

为了计算方便,若损失函数选择二次函数 $l(y,$

$\hat{y}) = (y - \hat{y})^2$,则有

$$h^* = \min_{h \in H} \frac{\lambda}{2} \|h\|_H^2 + \frac{1}{2} \sum_{i=1}^n (y_i - h(x_i))^2 = \frac{1}{2} \mathbf{y}^T \mathbf{B} \mathbf{y}$$

式中, $\mathbf{B} = (\mathbf{K} + \lambda \mathbf{I})^{-1}$, $\mathbf{K} = [\kappa(x_i, x_j)]_{n \times n}$ 为 $n \times n$ 的核矩阵。因此,评价函数 $\hat{f}(D_l, D_u, x_s)$ 可简化为

$$\hat{f}(D_l, D_u, x_s) = \min_{y_u \in \{-1, +1\}^{n_u-1}} \max_{y_s \in \{-1, +1\}} \mathbf{y}^T \mathbf{B} \mathbf{y} \quad (7)$$

为表示方便,在矩阵中用下标 u 表示 D_u 中的未标注样本, l 表示 D_l 中已标注的样本, s 表示当前选择的样本。下标 U 表示所有的未标注样本 ($D_u \cup \{x_s\}$)。基于这些约定, $\mathbf{y}^T \mathbf{B} \mathbf{y}$ 可重新表示为

$$\begin{aligned} \mathbf{y}^T \mathbf{B} \mathbf{y} &= \mathbf{y}_l^T \mathbf{B}_{ll} \mathbf{y}_l + \mathbf{B}_{ss} + \mathbf{y}_u^T \mathbf{B}_{uu} \mathbf{y}_u + \\ &+ 2\mathbf{y}_u^T (\mathbf{B}_{ul} \mathbf{y}_l + \mathbf{B}_{us} \mathbf{y}_s) + 2\mathbf{y}_s \mathbf{y}_l^T \mathbf{B}_{ls} \end{aligned}$$

注意到上面的目标函数对 y_s 是(线性)凹的,对 y_u 是二次凸的,可以利用式(7)将求 y_s 的最小值转化为求 y_u 的最大值。通过松弛 y_u 到连续变量,使 $\mathbf{y}^T \mathbf{B} \mathbf{y}$ 取最小值的 y_u 为

$$\hat{\mathbf{y}}_u = -\mathbf{B}_{uu}^{-1} (\mathbf{B}_{ul} \mathbf{y}_l + \mathbf{B}_{us} \mathbf{y}_s) \quad (8)$$

根据以上的描述,评价函数最终可表示为

$$\begin{aligned} \hat{f}(D_l, D_u, x_s) &= \mathbf{B}_{ss} + \mathbf{y}_l^T \mathbf{B}_{ll} \mathbf{y}_l + \max_{y_s = \pm 1} \{2\mathbf{y}_s \mathbf{B}_{sl} \mathbf{y}_l + \\ &+ (\mathbf{B}_{ul} \mathbf{y}_l + \mathbf{B}_{us} \mathbf{y}_s)^T \mathbf{B}_{uu}^{-1} (\mathbf{B}_{ul} \mathbf{y}_l + \mathbf{B}_{us} \mathbf{y}_s)\} \approx \\ &= \mathbf{B}_{ss} - \frac{\det(\mathbf{B}_{UU})}{\mathbf{B}_{ss}} + 2|(\mathbf{B}_{sl} - \mathbf{B}_{su} \mathbf{B}_{uu}^{-1} \mathbf{B}_{ul}) \mathbf{y}_l| \quad (9) \end{aligned}$$

式中,最后一步满足关系

$$\det \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \det(A_{22}) \det(A_{11} - A_{12} A_{22}^{-1} A_{21})$$

从式(9)可以看出,评价函数 $\hat{f}(D_l, D_u, x_s)$ 实质上由两部分组成: $\mathbf{B}_{ss} - \det(\mathbf{B}_{UU})/\mathbf{B}_{ss}$ 和 $|(\mathbf{B}_{sl} - \mathbf{B}_{su} \mathbf{B}_{uu}^{-1} \mathbf{B}_{ul}) \mathbf{y}_l|$ 。因 $\mathbf{B}_{UU}$ 独立于被选择的样本 x_s ,故最小化式(9)中的第1项等价于最小化 $\mathbf{B}_{ss}$ 。

因 $\mathbf{B} = (\mathbf{K} + \lambda \mathbf{I})^{-1}$,有

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} K_{ss} - (K_{sl}, K_{su}) \begin{pmatrix} K_{ll} & K_{lu} \\ K_{ul} & K_{uu} \end{pmatrix}^{-1} \begin{pmatrix} K_{ls} \\ K_{us} \end{pmatrix} \\ \frac{1}{K_{ss}} \left[1 + \frac{1}{K_{ss}} (K_{sl}, K_{su}) \begin{pmatrix} K_{ll} & K_{lu} \\ K_{ul} & K_{uu} \end{pmatrix}^{-1} \begin{pmatrix} K_{ls} \\ K_{us} \end{pmatrix} \right] \end{bmatrix} \approx \end{aligned}$$

因此,要想选择一个使 $\mathbf{B}_{ss}$ 较小的样本,需要先选择一个自身相似性 K_{ss} 较大的样本。当 K_{ss} 是常量时,这一项将不会影响选择结果。为了分析式(9)中第2项的影响,将其近似等价于

$$\begin{aligned} 2|(\mathbf{B}_{sl} - \mathbf{B}_{su} \mathbf{B}_{uu}^{-1} \mathbf{B}_{ul}) \mathbf{y}_l| &\approx 2|\mathbf{B}_{sl} \mathbf{y}_l| + \\ &+ 2|\mathbf{B}_{su} \mathbf{B}_{uu}^{-1} \mathbf{B}_{ul} \mathbf{y}_l| \approx \end{aligned}$$

$$2|B_{sl}y_l| + 2|B_{su}\hat{y}_u| \quad (10)$$

式中,第1项衡量了仅用已标注数据预测的 x_s 的置信度,对应于样本的信息量;第2项衡量了利用未标注样本的预测标签对 x_s 的置信度,可以视为样本的代表性。最终,利用 D_u 中的未标注数据预测 x_s 的标签,其结果依赖于类别标签 $\hat{y}_u$ 的平均值。在现有的算法中,衡量一个样本的代表性往往仅利用了未标注数据的聚类结构。而本文衡量样本的代表性取决于 y_u ,实质上是将未标注样本的聚类结构和已标注样本的类别标签结合起来。

2.2 AIR_OML 算法

一方面,基于 min-max 理论,QUIRE^[3]算法查询最具信息量和最具代表性的未标注样本,经人工标注后由当前的分类器利用进行再学习。另一方面,分类器再训练时应避免使用所有的训练数据进行重训练。并且,当新数据到来时,分类器应实现实时更新。这里,采用了文献[12]中提出的在线学习算法。结合 QUIRE 算法和在线学习算法,本文提出了结合主动学习的多标签图像在线分类算法,即 AIR-OML 算法。AIR_OML 算法的流程框图,如图 1 所示。主动学习挑选样本的 AL(active learn)算法步骤如算法 1,在线更新的 AIR-OML 算法步骤如算法 2。

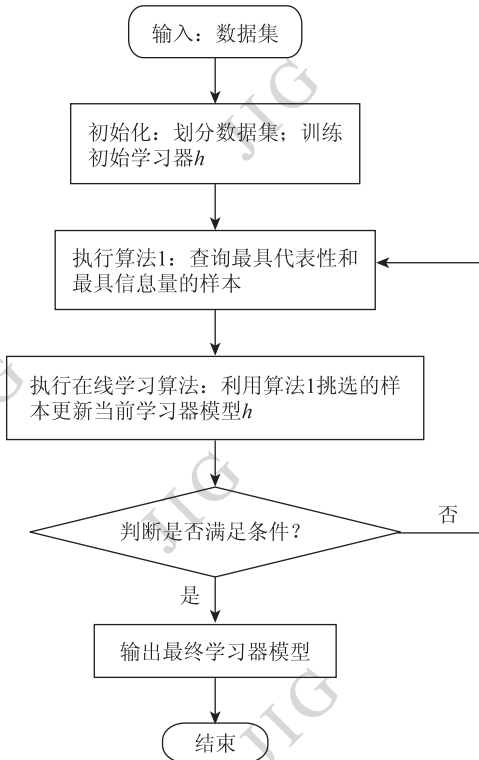


图1 AIR-OML 算法流程框图

Fig. 1 Flowchart of AIR-OML

算法1 AL 算法(选择样本)

输入:数据集 D ;

输出:样本-标签对 (x_s^*, y_s^*) ;

初始化。初始化数据 D_l, n_l, D_u, n_u ;

计算核函数 K ;

REPEAT

由 $\det(L_{UU})$ 计算 B_{UU}^{-1} ;

FOR $i = 1 : n_u$ DO

计算 B_{uu}^{-1}

根据式(9),计算 $\hat{f}(d_l, d_u, x_s)$

END FOR

选择 x_s^* 使得 $\hat{f}(d_l, d_u, x_s)$ 最小并查询其标签直到满足查询数目或是达到分类器所需精度要求。

END REPEAT

算法2 AIR-OML 算法

输入:数据集 X ;

输出:最终的学习模型;

1)初始化

将数据集 X 随机地划分为3个部分:已标注训练样本集 X_l 、未标注训练样本集 X_u 和测试样本集 X_t ,然后在已标注训练样本集上训练学习器 h ;

REPEAT

2)利用算法1查询最具代表性和最具信息量的样本,并进行标注;

3)更新当前学习模型

利用在线学习算法和步骤2)中选择的样本及其标签更新当前学习器模型;

4)重复步骤2),直到满足分类器精度要求或查询样本达到足够数目结束算法。

END REPEAT

3 实验

实验中采用 Image、yeast、CAL500 和 Emotions (相关描述见表1)数据集,利用4种常用的多标签学习评价指标汉明损失(Hamming Loss)、排序损失(Ranking Loss)、平均准确率(Average Precision)和覆盖率(Coverage)对算法性能进行评价。评价指标的详细描述见文献[1],这里不再赘述。在此,采用的样本选择策略与随机挑选样本方法和基于间隔的样本方法^[16]进行了对比。基准分类器选用 Rank-SVM^[19],支持向量机(SVM)中的核函数这里采用多

项式。每个数据集按比例被随机地分成 3 个部分：已标注数据（少量）、未标注数据（大量）和测试数据。在初始阶段，随机选择样本用于训练初始分类器模型，每一次迭代中只选择一个样本请求标注，实验中每查询 10 个样本或 20 个样本对测试样本集预测一次。另外，图 2 统计了各数据集中标签的分布信息。

表 1 数据集描述

Table 1 Description of 4 datasets

数据集	样本数	标签数	特征数
CAL500	502	174	68
Emotions	593	6	72
Image	2 000	5	294
Yeast	2 417	14	103

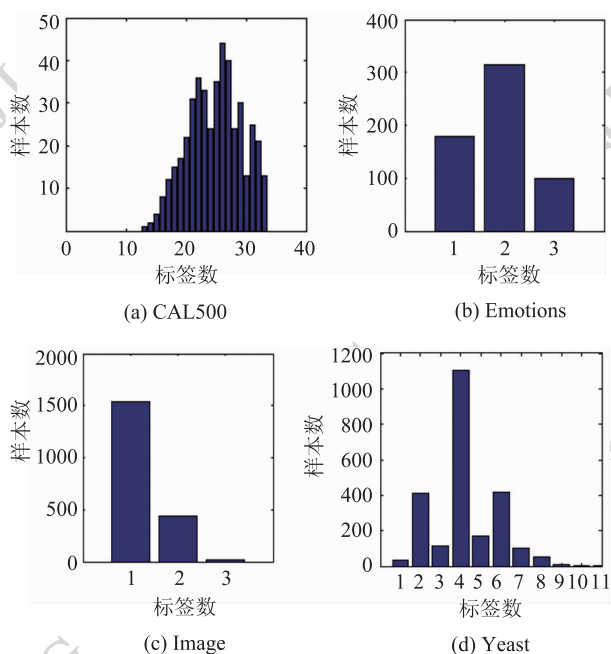


图 2 数据集上的样本标签数目分布信息

Fig. 2 Distribution of label numbers for samples on four datasets

根据以上设置，随着查询样本的增加，分别画出在 4 个数据集上在 4 种度量标准下的实验结果对比曲线，如图 3—图 6 所示。其中，I&R 代表本文的挑选样本方法，RANDOM 代表随机挑选样本方法，MARGIN 是基于间隔的样本挑选方法。从图 3—图 6 中可以看出，基于查询最具代表性和最具信息量样本的选择策略，随着训练集中样本的不断加，分类器的性能也得到不断改善。

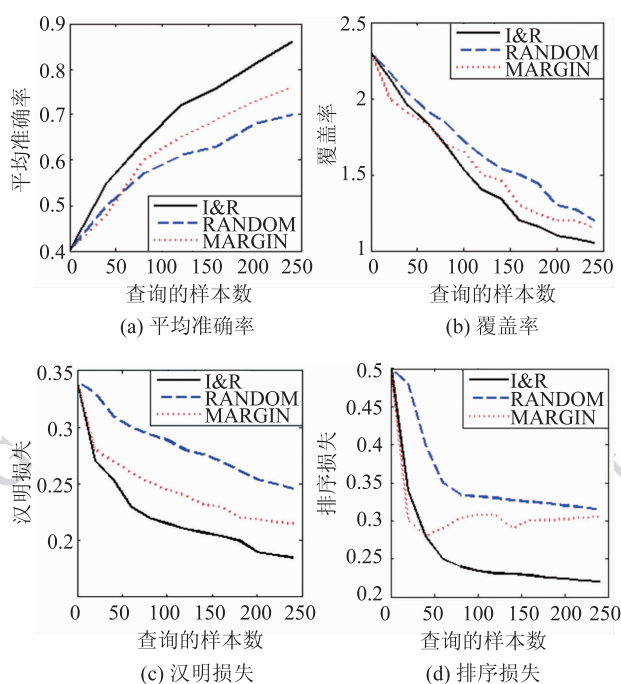


图 3 数据集 CAL500 上 4 种评价指标的对比结果

Fig. 3 Comparisons on CAL500 with four evaluation metrics

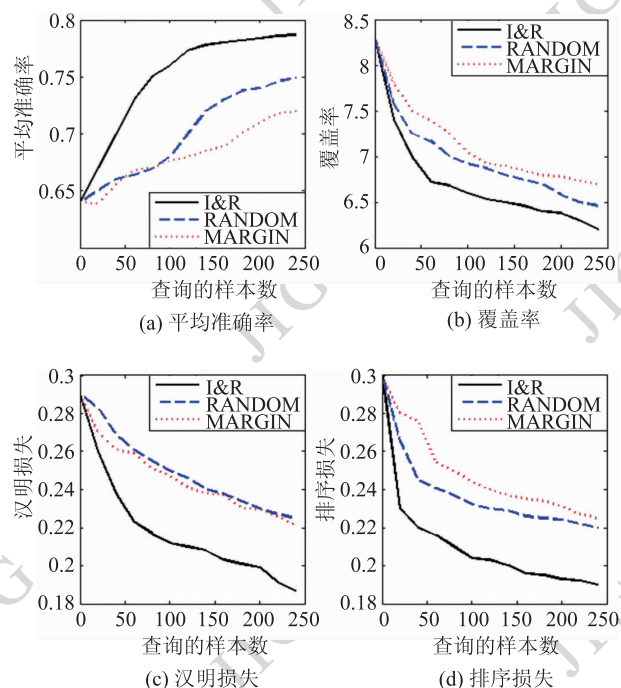


图 4 数据集 Emotions 上 4 种评价指标的对比结果

Fig. 4 Comparisons on Emotions with four evaluation metrics

另外，与已有的样本挑选方法：随机挑选方法和基于间隔的样本挑选方法相比，本文的查询最具代表性和最具信息量的样本挑选方法在 4 种评价指标上均占据明显优势。同时，还可以看出，当分类器达到相同或相近的分类准确度时，本文方法所选择的

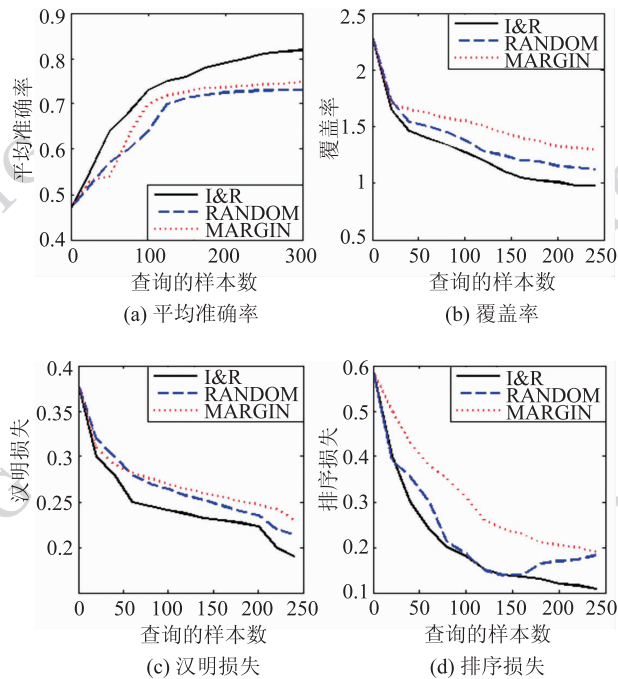


图5 数据集 Image 上 4 种评价指标的对比结果

Fig. 5 Comparisons on Image with four evaluation metrics

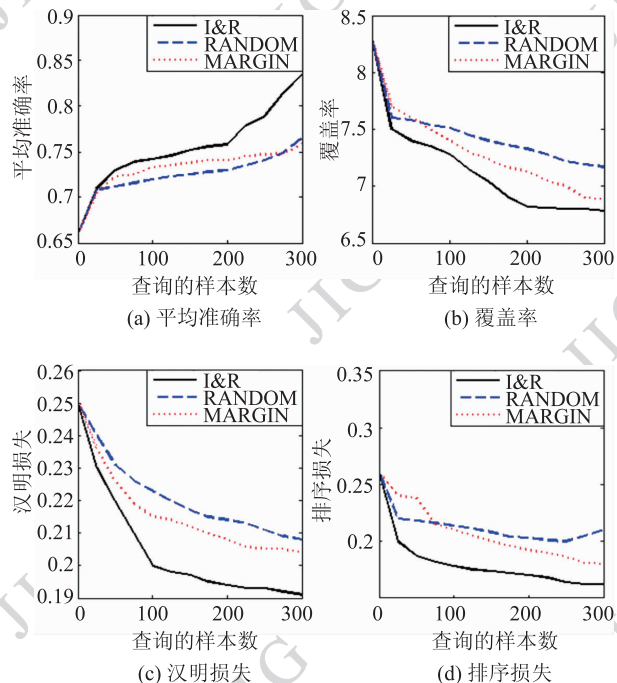


图6 数据集 Yeast 上 4 种评价指标的对比结果

Fig. 6 Comparisons on Yeast with four evaluation metrics

样本数目要明显少于采用随机挑选样本方法和基于间隔的采样方法所需查询的样本数。与此同时,实验结果也表明采用本文的样本选择方法在保证分类器性能的前提下,使用尽可能少的已标注训练样本,可以最大限度地降低人工标注样本的代价。

再者,在主动学习过程中,随着查询样本的增加,训练样本的规模不断扩大,分类器相应地也需要及时更新。通常,传统的分类器再训练需要使用全部的训练样本进行学习。使用大量的训练样本进行分类器学习,不仅费时巨大,也需要消耗大量的存储空间。为解决此难题,本文采用一种高效的在线学习方法,它不需要用到既往的样本,只需要利用新标注的样本对现有的分类模型进行更新。为了说明该算法的有效性,在此将该算法与 SVM 重训练方法在每次更新模型所需的平均计算时间进行了验证对比(4 核计算机,CPU 主频 1.8 GHz,内存 16 G),对比结果如表 2。从表 2 中可以看出,在线学习在模型学习方面要比 SVM 重训练快近一个数量级。由于本文用的数据集规模相对较小,对比效果不是很明显。当使用较大规模的数据集时,在线学习更新模型的效率将表现更突出。

表2 更新模型所用的平均计算时间

Table 2 Average computing time for updating/retraining models

数据集	训练模型	计算时间/min
CAL500	在线学习	5.92
	SVM 重训练	63.3
Emotions	在线学习	2.2
	SVM 重训练	15.46
Image	在线学习	1.35
	SVM 重训练	10.26
Yeast	在线学习	4.57
	SVM 重训练	31.87

4 结 论

本文目的是在多标签学习框架中,减少人工标注代价,提高分类器再学习效率,从而构建一个具有较强泛化性能和较快训练效率的多标签图像分类器。针对初始训练样本数量较少的问题,采用主动学习方法查询最具代表性和最具信息量的样本来增加训练样本数量。利用主动学习方法,人工标注样本的代价得到显著降低。另外,针对传统的分类器重训练效率低下的问题,基于 KKT 条件通过度量新

旧分类器之间的距离,利用新标注的样本实现已有分类器的在线更新,克服了传统分类器再训练更新效率低下的问题,大大提高了分类器再学习的效率,并有效地解决了训练样本存储空间负担的问题。在 Image、Yeast、CAL500 和 Emotions 4 个公开数据集上的实验结果验证了本文方法的有效性。在未来工作中,拟引入标签之间的关联关系^[20],以减少信息冗余,进一步提高算法的分类效果。

参考文献 (References)

- [1] Schapire R E, Singer Y. BoostTexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39 (2-3): 135-168. [DOI: 10.1023/A:1007649029923]
- [2] Hoi S C H, Jin R, Zhu J K, et al. Semi-supervised svm batch mode active learning for image retrieval [C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Anchorage, AK: IEEE, 2008: 1-7. [DOI: 10.1109/CVPR.2008.4587350]
- [3] Huang S J, Jin R, Zhou Z H. Active learning by querying informative and representative examples [C]//Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: NIPS, 2010: 892-900. [DOI:10.1109/TPAMI.2014.2307881]
- [4] Balcan M F, Broder A Z, Zhang T. Margin based active learning [C]// Proceedings of the 20th Annual Conference on Learning Theory. San Diego, CA, USA: Springer Berlin Heidelberg, 2007: 35-50.
- [5] Freund Y, Seung H S, Shamir E, et al. Selective sampling using the query by committee algorithm [J]. Journal of Machine Learning, 1997, 28(2-3): 133-168. [DOI: 10.1023/A:1007330508534]
- [6] Yu K, Bi J, Tresp V. Active learning via transductive experimental design [C]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, USA: ACM, 2006: 1081-1088. [DOI: 10.1145/1143844.1143980]
- [7] Dasgupta S, Hsu D. Hierarchical sampling for active learning [C]//Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008: 208-215.
- [8] Donmws P, Carbonell J G, Bennett P N. Dual strategy active learning [C]//Proceedings of the 18th European Conference on Machine Learning. Berlin: Springer, 2007: 116-127. [DOI: 10.1007/978-3-540-74958-5_14]
- [9] Nguyen H T, Smeulder A W. Active learning using pre-clustering [C]//Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004: 623-630.
- [10] Li X C, Wang L, Sung E. Multi-label SVM active learning for image Classification [C]// Proceedings of IEEE International Conference on Image Processing. Singapore: University of Nan-yang Technology, 2004: 2207-2210. [DOI: 10.1109/ICIP.2004.1421535]
- [11] Qi G J. Multi-label classification with its application into multimedia analysis [D]. Hefei: University of Science and Technology of China, 2009. [齐国君. 多类别模式分类技术及其在多媒体分析上的应用 [D]. 合肥: 中国科学技术大学, 2009.]
- [12] Qi G J, Hua X S, Rui Y, et al. Two-dimensional multi-label active learning with an efficient online adaption model for image classification [J]. IEEE Transactions on Pattern analysis and Machine Intelligence, 2009, 31 (10): 1880-1897. [DOI: 10.1109/TPAMI.2008.218]
- [13] Jiang H, Qi Y S. Active learning for multi-label classification based on sphere structured support vector machine [J]. Journal of Computer Applications, 2012, 32(5): 1359-1366. [蒋华, 戚玉顺. 基于球结构支持向量机的多标签分类的主动学习 [J]. 计算机应用, 2012, 32(5): 1359-1366.]
- [14] Zhang X Y, Cheng J, Xu C S, et al. Multi-view multi-label active learning for image classification [C]//Proceedings of IEEE International Conference on Multimedia and Expo. New York: IEEE, 2009: 258-261. [DOI: 10.1109/ICME.2009.5202484]
- [15] Brinker K. On Active Learning in Multi-label Classification [M]. Berlin Heidelberg: Springer, 2006: 206-213. [DOI: 10.1007/3-540-31314-1_24]
- [16] Liu D Y, Qiu W J. Active learning method for multi-label classification based on average expectation margin [J]. Journal of Computer Engineering, 2011, 37(15): 168-170. [刘端阳, 邱卫杰. 基于平均期望间隔的多标签分类主动学习方法 [J]. 计算机工程, 2011, 37(15): 168-170.]
- [17] Xu Y. Multi-label SVM active learning based on one side of margin [J]. Journal of Computer Knowledge and Technology, 2012, 8(35): 8349-8352. [许钰. 基于偏侧分类间隔的多标签 SVM 主动学习 [J]. 电脑知识与应用, 2012, 8(35): 8349-8352.]
- [18] Belkin M, Niyogi P, Sindwani V. Manifold regularization: a geometric framework for learning from labeled and by batch mode active learning [J]. Journal of Machine Learning Research, 2006, 1(7): 633-642. [DOI: 10.1145/1135777.1135870]
- [19] Elisseeff A, Weston J. A kernel method for multi-labeled classification [C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 681-687. [DOI:10.1109/2013.062613.00160]
- [20] Sun F M, Tang J H, Li H J, et al. Multi-label image categorization with sparse factor representation [J]. IEEE Transactions on Image Processing, 2014, 23(3): 1028-1037. [DOI: 10.1109/TIP.2014.2298978]