DOI:10.3969/j.issn.1673-4785.201305033
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130929.1105.006.html

基于 Tri-training 的半监督多标记学习算法

刘杨磊[1,2],梁吉业[1,2],高嘉伟[1,2],杨静[1,2]
(1.山西大学 计算机与信息技术学院,山西 太原 030006; 2.山西大学 计算智能与中文信息处理教育部重点实验室,山西 太原 030006)

摘　要:传统的多标记学习是监督意义下的学习,它要求获得完整的类别标记.但是当数据规模较大且类别数目较多时,获得完整类别标记的训练样本集是非常困难的.因而,在半监督协同训练思想的框架下,提出了基于 Tri-training 的半监督多标记学习算法(SMLT).在学习阶段,SMLT 引入一个虚拟类标记,然后针对每一对类别标记,利用协同训练机制 Tri-training 算法训练得到对应的分类器;在预测阶段,给定一个新的样本,将其代入上述所得的分类器中,根据类别标记得票数的多少将多标记学习问题转化为标记排序问题,并将虚拟类标记的得票数作为阈值对标记排序结果进行划分.在 UCI 中 4 个常用的多标记数据集上的对比实验表明,SMLT 算法在 4 个评价指标上的性能大多优于其他对比算法,验证了该算法的有效性.

关键词:多标记学习;半监督学习;Tri-training
中图分类号:TP181　**文献标志码**:A　**文章编号**:1673-4785(2013)05-439-07

中文引用格式:刘杨磊,梁吉业,高嘉伟,等.基于 Tri-training 的半监督多标记学习算法[J]. 智能系统学报, 2013, 8(5): 439-445.
英文引用格式:LIU Yanglei, LIANG Jiye, GAO Jiawei, et al. Semi-supervised multi-label learning algorithm based on Tri-training [J]. CAAI Transactions on Intelligent Systems, 2013, 8(5): 439-445.

Semi-supervised multi-label learning algorithm based on Tri-training

LIU Yanglei[1,2], LIANG Jiye[1,2], GAO Jiawei[1,2], YANG Jing[1,2]
(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: Traditional multi-label learning is in the sense of supervision, in which the complete category labels are required. However, when the size of data is large and there are several categories of labels, it is quite difficult to obtain the training sample sets with complete labels. Therefore, a semi-supervised multi-label learning algorithm based on Tri-training (SMLT) is proposed. In the learning stage, SMLT initially introduces a virtual label, then for each pair of virtual labels, the Tri-training algorithm is utilized to train the corresponding classifiers for each pair of labels. In the forecast stage, a new sample is given, which will be substituted into the obtained classifier described above. According to the votes of each label, the multi-label learning problem is transformed into a label ranking problem, subsequently; the votes of the virtual label are taken as the threshold for distinguishing the label ranking results. The contrast experiments on four commonly used UCI multi-label datasets show the SMLT algorithm behaves better than other comparative algorithms in four evaluation indices and the effectiveness of the proposed algorithm is verified.

Keywords: multi-label learning; semi-supervised learning; Tri-training

多标记学习(multi-label learning)[1]是机器学习领域的重要研究方向之一.在多标记学习问题中,一个训练样本可能同时对应于一个或多个不同的概念标记,以表达其语义信息,学习的任务是为待学习样本预测其对应的概念标记集合.多标记学习问题普遍存在于真实世界中,比如在图像场景分类任务中,一幅图像可能因包含“树木”、“天空”、“湖泊”以及“山峰”等语义概念,而拥有多个概念标记.

收稿日期:2013-05-09.　**网络出版日期**:2013-09-29.
基金项目:国家“973”计划前期研究专项(2011CB311805);山西省科技攻关计划资助项目(20110321027-01);山西省科技基础条件平台建设项目(2012091002-0101).
通信作者:梁吉业. E-mail: ljy@ sxu.edu.cn.

传统的多标记学习通常是在监督意义下进行的,即要求训练数据集的训练样本必须全部是已标记样本.然而,在现实生活中,虽然获取大量的训练数据集并不十分困难,但是为这些数据提供正确的类别标记却需要耗费大量的人力和时间.比如,在图像场景分类任务中,现实世界中存在着海量的未标记图像,而且一幅图像往往拥有大量的候选类别标记,要完整标注训练集中的每一个样本就意味着需要人工查看每一幅图像的所有候选类别并逐一标注.当数据规模较大且类别数目较多时,要获得完整类别标记的训练样本集是非常困难的.此时,在监督意义下如果只使用少量已标记样本训练,则得到的模型很难具有较强的泛化能力.而半监督学习能够较好地解决上述问题,它综合利用少量的已标记样本和大量的未标记样本以提高泛化性能[2-3].

因此,本文主要以协同训练思想为核心,提出了基于 Tri-training 的半监督多标记学习算法(a semi-supervised multi-label learning algorithm based on Tri-training, SMLT),以解决广泛存在于实际生活中的文本分类、图像场景分类以及生物信息学等半监督多标记学习问题.

1 背景知识

1.1 多标记学习

在多标记学习框架下,每个对象由一个样本描述,该样本具有多个类别标记,学习的目的是将所有合适的类别标记赋予待学习样本[4].形式化地来说,令 X 表示样本空间,Y 表示类别标记空间,给定数据集 $\{(x_1,Y_1),(x_2,Y_2),\cdots,(x_m,Y_m)\}$,目标是学得 $f:X\to 2^Y$.其中,$x_i\in X(i=1,2,\cdots,m)$ 为一个样本,$Y_i\subseteq Y$ 为 x_i 的一组类别标记 $\{y_{i1},y_{i2},\cdots,y_{in}\}$,$y_{ij}\in Y(j=1,2,\cdots,n)$,$n$ 为 Y_i 中所含类别标记的个数.

如果限定每个样本只对应一个类别标记,那么传统的 2 类或多类学习问题均可以看作是多标记学习问题的特例.然而,多标记学习的一般性也使得其相较于传统的学习问题更加难以解决.目前,多标记学习面临的最大挑战在于其输出空间过大,即与一个待学习样本相关联的候选类别标记集合的数量将会随着标记空间的增大而呈指数规模增长.如何充分利用标记之间的相关性是构造具有强泛化能力多标记学习系统的关键.根据考察标记之间相关性的不同方式,已有的多标记学习问题求解策略大致可以分为以下 3 类[5]:

1)"一阶"策略:将多标记学习问题分解为多个独立的二分类问题进行求解.该类方法效率较高且实现简单,但是由于忽略了标记之间的相关性,通常学习系统的泛化性能较低.

2)"二阶"策略:考察两两标记之间的相关性,将多标记学习问题转化成标记排序问题进行求解.该类方法在一定程度上考虑了标记之间的相关性,学习系统的泛化性能较好,但是当实际问题中标记之间具有超越二阶的相关性时,该类方法的性能将会受到很大影响.

3)"高阶"策略:考察高阶的标记相关性,充分利用标记之间的结构信息进行求解.该类方法可以较好地反映真实世界问题的标记相关性,但其模型复杂度较高,且在缺乏领域知识指导的情况下,几乎无法利用标记之间的结构信息.

另一方面,近几年来,多标记学习越来越受到机器学习领域学者的关注,研究人员对多标记学习问题提出了许多学习方法和策略,对这些问题的研究大致可分为 2 种思路:一种是问题转化,另一种是算法改编.第 1 种思路试图将多标记学习任务转化为一个或多个单标记学习任务,然后利用已有的学习算法求解.代表性学习算法有一阶方法 Binary Relevance[6],它将多标记学习问题转化为二分类问题进行求解;二阶方法 Calibrated Label Ranking[7]将多标记学习问题转化为标记排序问题求解;高阶方法 Random k-labelsets[8]将多标记学习问题转化为多类分类问题求解.第 2 种思路是对现有算法进行改编或设计新算法,使之能直接处理多标记学习任务.代表性学习算法有一阶方法 ML-kNN[9],它将"惰性学习"算法 k 近邻进行改造以适应多标记数据;二阶方法 Rank-SVM[10]将"核学习"算法 SVM 进行改造用于多类别标记;高阶方法 LEAD[5]将"贝叶斯学习"算法中的 Bayes 网络进行改造,以适应多标记数据.

上述的多标记学习算法通常为监督学习算法.然而,为训练数据集提供正确的类别标记需要耗费大量的人力和时间.因此,当只有少量已标记样本可用时,传统的多标记学习算法将不再适用.

1.2 半监督多标记学习

近来年,有一些研究者开始研究半监督/直推式多标记学习(semi-supervised/transductive multi-label learning)方法.半监督学习和直推式学习都是试图利用大量的未标记样本来辅助对少量已标记样本的学习,但二者的基本思想却有显著的不同.直推式学习的测试样本是训练集中的未标记样本,测试环境是封闭的;而半监督学习的测试样本与训练样本无关,测试环境是相对开放的.

2006 年,Liu 等[11]基于如果样本之间具有很大的相似性,那么它们的标记集合之间也应该具有很

大的相似性这样的假设,提出了 CNMF(constrained non-negative matrix factorization)方法,通过解一个带约束的非负矩阵分解问题,期望使得这 2 种相似性差值最小,从而获得最优的对未标记样本的标记.2008 年,姜远等[12]提出了基于随机游走(random walk)的直推式多标记学习算法 TML,并将其用于文本分类.同年,Chen 等[13]基于样本相似性度量与标记相似性度量构建图,提出了 SMSE(semi-supervised algorithm for multi-label learning by solving a Sylvester equation)方法,采用标记传播的思想对未标记样本的标记进行学习,整个优化问题可采用 Sylvester 方程进行快速求解.2010 年,Sun 等[14]和周志华等[15]考虑多标记学习中的弱标记问题,即训练样本对应的标记集合中只有一小部分得到了标记,或者根本没有任何的标记,分别提出了 WELL(weak label learning)方法和 TML-WL(transductive multi-label learning method for weak labeling)方法,他们同样采用标记传播的思想对缺失标记进行学习.2013 年,周志华等[16]还采用标记传播的思想,首先将学习任务看作是一个对标记集合进行估计的优化问题,然后为这个优化问题找到一个封闭解,提出的 TRAM 算法为未标记样本分配其对应的标记集合.以上方法都是直推式方法,这类方法不能自然地对除测试样本以外的未见样本进行预测.2012 年,周志华等[17]在传统经验风险最小化原理基础上,引入 2 种正则项分别用于约束分类器的复杂度和相似样本拥有相似结构化的多标记输出,针对归纳式半监督多标记学习,提出了一种正则化方法 MASS(multi-label semi-supervised learning).

1.3　Tri-training 算法

从 20 世纪 90 年代末标准协同训练算法被提出开始,很多研究者对协同训练技术进行了研究,不仅提出了很多学习方式不同、限制条件强弱各异的算法,而且对协同训练的理论分析和应用研究也取得了不少进展,使得协同训练成为半监督学习中重要的研究方向之一[18].

初期的协同训练算法引入了很多的限制和约束条件.而 Tri-training 算法[19]是周志华等在 2005 年提出的一种新的协同训练方法,它使用 3 个分类器进行训练.在学习过程中,Tri-training 算法采用集成学习中经常用到的投票法,使用 3 个分类器对未见样本进行预测.

由于 Tri-training 对属性集和 3 个分类器所用监督学习算法都没有约束,而且不使用交叉验证,其适用范围更广、效率更高,因此本文以协同训练思想为核心,利用 Tri-training 算法训练分类器,来研究半监督多标记学习.

2　基于 Tri-training 的半监督多标记学习算法

下面提出一种基于 Tri-training 的半监督多标记学习算法,该算法考察两两标记之间的相关性,将多标记学习问题转化为标记排序问题进行求解;因此在一定程度上考虑了标记之间的相关性,并采用半监督学习中的协同训练思想,利用 Tri-training 过程来训练分类器.

本文中相关量的定义如下:$L=\{(x_i,Y_i),i=1,2,\cdots,m\}$ 是包含 m 个样本的已标记样本集.其中,x_i 表示第 i 个样本的属性集合;$Y_i=\{y_{i1},y_{i2},\cdots,y_{in}\}$ 表示样本 x_i 对应的包含 n 个标记的类别标记集合,且 $y_{ij}\in\{0,1\}$,$j=1,2,\cdots,n$,若 $y_{ij}=1$,则表示第 j 个标记是当前样本 x_i 的真实标记,否则 $y_{ij}=0$. $U=\{x_k',k=1,2,\cdots,t\}$ 是包含 t 个样本的未标记样本集.$L\cup U$ 是包含 $m+t$ 个样本的训练集.为了验证所提分类算法的有效性,构建的 $T=\{x_s'',s=1,2,\cdots,w\}$ 是包含 w 个样本的测试集.数组 $R_{sj}(s=1,2,\cdots,w,j=1,2,\cdots,n)$ 用于存放测试集 T 中样本 x_s''在第 j 类标记上的得票数.

为了对后续过程中产生的标记排序结果进行分析,并得到最终的预测标记集合,需要设置一个阈值来划分上述标记排序结果.因此,在算法的预处理阶段,为每一个训练样本 x_i 添加一个虚拟标记 y_{i0},把虚拟类标记的得票数作为阈值对标记排序结果进行划分.此时,涉及到标记的下标应从 0 开始.

SMLT 算法的基本思想是:首先,为已标记样本集 L 中的每一个样本 x_i 添加一个虚拟标记 y_{i0},然后考虑两两标记之间的相关性,对 L 中每一对标记 $(y_{*p},y_{*q})(0\leqslant p<q\leqslant n)$ 进行训练,并利用 Tri-training 过程学习得到相应的 3 个分类器.对一个新的测试样本,用学习到的分类器对相应的每一对标记进行预测,并统计每个标记所得的票数 R_{sj},得到该测试样本的一个标记排序结果.最后以虚拟标记 y_{s0}''的得票数 R_{s0}作为确定类标记的依据,若 $R_{sj}>R_{s0}(j=1,2,\cdots,n)$,则样本 x_s''的最后标记 $y_{sj}''=1$,否则 $y_{sj}''=0$,即可得到一组测试集样本的预测结果 Y''.

SMLT 算法的流程如图 1 所示.SMLT 算法的详细步骤如下.

输入:已标记样本集 L,未标记样本集 U,测试集 T.

输出:对测试集 T 的预测结果 Y''.

1)初始化 $R_{sj}=0(s=1,2,\cdots,w,j=0,1,\cdots,n)$ 和用于存放训练样本的集合 $V_{pq}=\varnothing(0\leqslant p<q\leqslant n)$.

2)预处理已标记样本集 L.对于任一对未处理的标记(y_{*p},y_{*q}),遍历 $x_i \in L$,将满足以下规则的 x_i 放入集合 V_{pq} 中.若 $y_{ip}=1,y_{iq}=0$ 则样本$(x_i,1)$放入集合 V_{pq} 中;若 $y_{ip}=0,y_{iq}=1$ 则将样本$(x_i,0)$放入集合 V_{pq} 中;若 $y_{ip}=y_{iq}$ 则不考虑样本 x_i,即样本 x_i 不放入集合 V_{pq} 中.

3)将集合 V_{pq} 作为新的已标记样本集 L^{new},结合未标记样本集 U,在训练集中利用 Tri-training 算法学习得到 3 个分类器.

4)使用投票法和得到的 3 个分类器对测试集 T 中的未标记样本 $x_s''(s=1,2,\cdots,w)$ 进行预测,得到预测结果 r_{spq} 并统计对应的标记投票个数.若 $r_{spq}=1$ 则表示样本 x_s'' 属于第 p 类标记,$R_{sp}=R_{sp}+1$;若 $r_{spq}=0$ 则表示样本 x_s'' 属于第 q 类标记,$R_{sq}=R_{sq}+1$.

5)将标记(y_{*p},y_{*q})设置为已处理,若还有未处理的标记对,则转步骤 2),否则下一步.

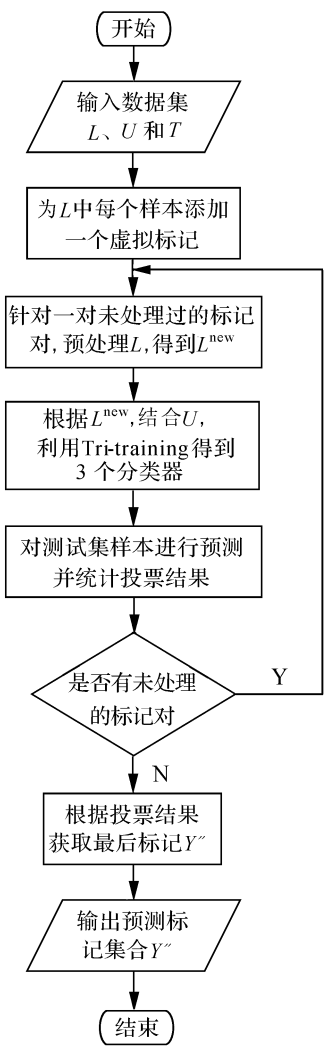

图 1 SMLT 算法

Fig.1 Flow chart of the SMLT algorithm

6)对于测试集 T 中的未标记样本 x_s'',若 $R_{sj}>R_{s0}(j=1,2,\cdots,n)$,则样本 x_s'' 的最后标记 $y_{sj}''=1$,否则 $y_{sj}''=0$,最终输出预测标记集合 $Y''=\{Y_s'',s=1,2,\cdots,w\}$.

3 实验结果及分析

本文在 emotions、scene、yeast、enron 这 4 个较为常用的多标记数据集[20]上与多标记学习的多种典型方法进行实验比较,其中包括 ML-kNN[9]、RANK-SVM[10]以及 TRAM[16].实验数据集的相关信息如表 1 所示.

表 1 实验数据集相关信息

Table 1 The characteristics of datasets

数据集名称	所属领域	样本个数	属性个数	类标记个数
emotions	music	593	72	6
scene	image	2 407	294	6
yeast	biology	2 417	103	14
enron	text	1 702	1 001	53

实验采用 4 种常用的多标记学习评价指标[4]对算法性能进行评估:Hamming Loss、One-Error、Coverage 和 Ranking Loss.以上 4 种评估指标的值越小,表明该算法的性能越好[4].

实验将抽取各数据集的 90% 作为训练样本集(其中 20% 的训练样本是已标记样本集,80% 的训练样本是未标记样本集),其余 10% 的数据为测试样本集,重复 10 次统计其平均结果.由于 TRAM 方法是直推式方法,不能直接对测试样本集以外的未见样本进行预测,实验中将最终测试样本作为 TRAM 训练时的未标记样本集.表 2~5 给出了实验结果,加粗部分为每个指标上的最佳性能.

表 2 数据集 emotions 上各算法的实验结果

Table 2 The summary results of four algorithms on emotions dataset

算法	Hamming Loss	One-Error	Coverage	Ranking Loss
MLkNN	0.257 1	0.406 8	2.203 4	0.239 9
RankSVM	0.279 7	0.423 7	2.237 3	0.278 1
TRAM	0.276 8	0.339 0	2.152 5	0.232 1
SMLT	**0.242 0**	**0.313 9**	**1.797 0**	**0.184 5**

表 3　数据集 scene 上各算法的实验结果

Table 3　The summary results of four algorithms on scene dataset

算法	Hamming Loss	One-Error	Coverage	Ranking Loss
MLkNN	**0.098 9**	0.253 1	0.560 2	0.095 5
RankSVM	0.112 7	0.232 4	0.473 0	**0.076 8**
TRAM	0.101 0	0.269 7	0.510 4	0.085 4
SMLT	0.114 1	**0.217 8**	**0.459 6**	0.077 1

表 4　数据集 yeast 上各算法的实验结果

Table 4　The summary results of four algorithms on yeast dataset

算法	Hamming Loss	One-Error	Coverage	Ranking Loss
MLkNN	**0.204 3**	0.235 6	6.425 6	0.173 3
RankSVM	0.208 4	0.219 0	6.388 4	0.177 8
TRAM	0.221 4	0.334 7	6.500 0	0.187 9
SMLT	0.210 5	**0.217 2**	**6.316 8**	**0.168 1**

表 5　数据集 enron 上各算法的实验结果

Table 5　The summary results of four algorithms on enron dataset

算法	Hamming Loss	One-Error	Coverage	Ranking Loss
MLkNN	0.058 7	0.370 6	15.300 0	0.104 8
RankSVM	0.074 7	0.355 9	14.065 9	0.099 6
TRAM	0.053 3	0.241 2	13.852 9	0.087 5
SMLT	**0.048 8**	**0.164 7**	**13.652 8**	**0.085 7**

通过分析表 2~5,在 emotions 和 enron 这 2 个数据集上,提出的算法 SMLT 在 4 个评估指标上都优于其他算法,而在 scene 数据集上有 2 个评估指标优于其他算法,但在 Hamming Loss 和 Ranking loss 上略差于其他算法,在 yeast 数据集上有 3 个评估指标优于其他算法,仅在 Hamming Loss 上略差于其他算法.可能的原因是本文提出的算法充分利用了已标记样本集和未标记样本集的信息,这要比不利用已标记样本集或者单纯只利用已标记样本集的信息,更能提高分类算法的性能.

为了进一步验证已标记样本集的规模对 SMLT 算法的影响,在 4 个数据集上分别进行实验.训练样本集和测试样本集的构成方法与上文实验相同,但是已标记样本集占训练数据集的比例依次调整为 10%、20%、30%、40%和 50%时,SMLT 算法在 4 项评估指标上的取值与已标记样本集比例的关系如图 2~5 所示.

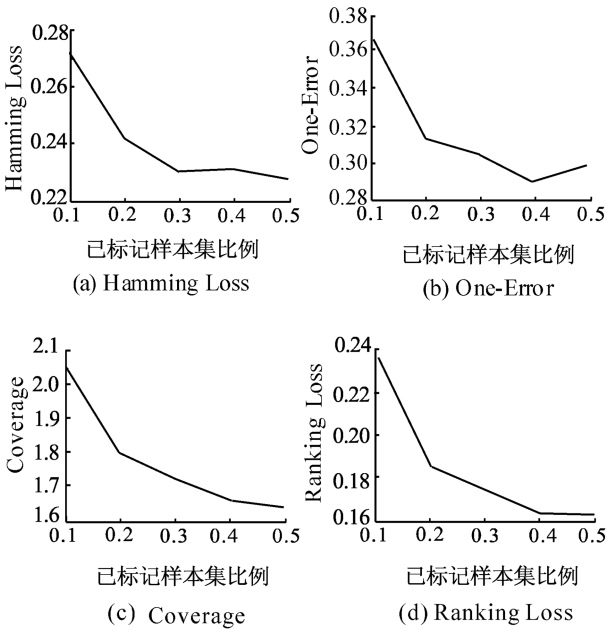

图 2　数据集 emotions 在 4 项评估指标上的实验结果

Fig.2　The summary results of four evaluation metrics on emotions dataset

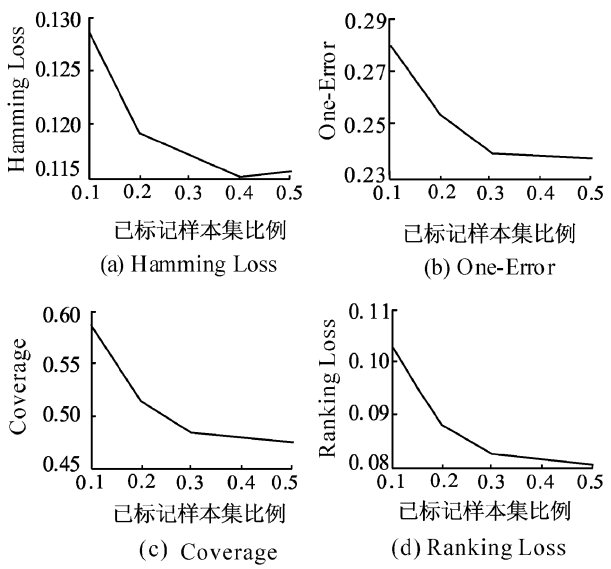

图 3　数据集 scene 在 4 项评估指标上的实验结果

Fig.3　The summary results of four evaluation metrics on scene dataset

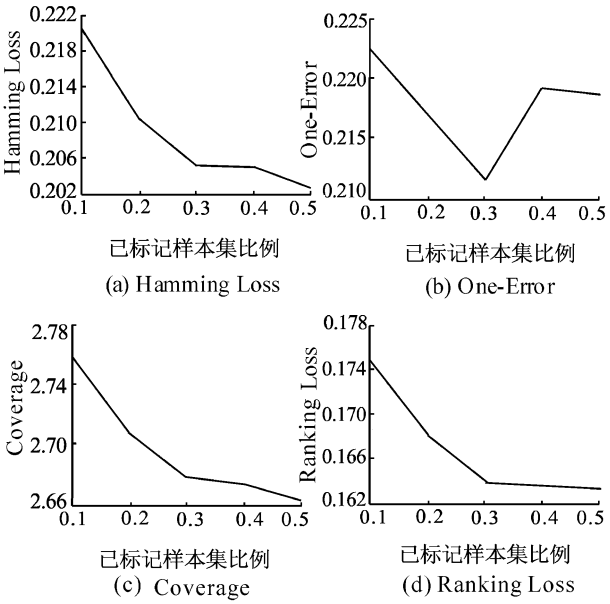

图 4 数据集 yeast 在 4 项评估指标上的实验结果

Fig.4 The summary results of four evaluation metrics on yeast dataset

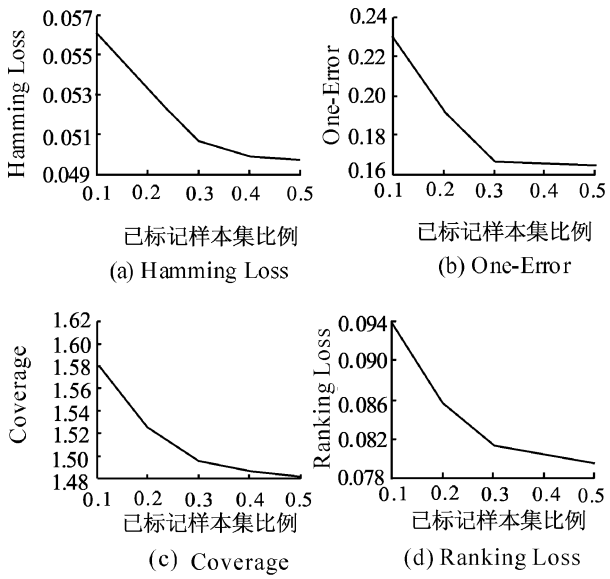

图 5 数据集 enron 在 4 项评估指标上的实验结果

Fig.5 The summary results of four evaluation metrics on enron dataset

根据图 2~5 可以发现,在半监督学习的意义下,SMLT 算法对应的 4 项评估指标的值大多随着已标记样本集比例的增加而不断减小,即算法的分类性能越来越好.并且在已标记样本集比例较小时曲线下降较快,随着已标记样本集比例的增加,曲线趋于平缓.仅在 yeast 数据集上的 One-Error 评价指标的曲线比较特殊.这是因为给定的监督信息越多,越有助于分类,从而得到更好的分类结果,而当已标记样本集比例增加到一定程度时,监督信息不再是影响分类性能的主要因素.

4 结束语

本文针对广泛存在于实际生活中的半监督多标记学习问题,以协同训练思想为核心,以两两标记之间的关系为出发点,利用 Tri-training 算法训练分类器,并将多标记学习问题转化为标记排序问题进行求解,实验结果表明了该算法的有效性.但是,当多标记的数量和规模较大时,如何进一步降低算法的计算复杂度仍将是需要深入讨论的问题.

参考文献:

[1] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13.

[2] ZHU Xiaojin. Semi-supervised learning literature survey [R]. Madison, USA: University of Wisconsin-Madison, 2008.

[3] 常瑜,梁吉业,高嘉伟,等.一种基于 Seeds 集和成对约束的半监督聚类算法[J].南京大学学报:自然科学版, 2012, 48(4): 405-411.
CHANG Yu, LIANG Jiye, GAO Jiawei, et al. A semi-supervised clustering algorithm based on seeds and pair wise constraints[J]. Journal of Nanjing University: Natural Sciences, 2012, 48(4): 405-411.

[4] ZHOU Zhihua, ZHANG Minling, HUANG Shengjun, et al. Multi-instance multi-label learning [J]. Artificial Intelligence, 2012, 176(1): 2291-2320.

[5] ZHANG Minling, ZHANG Kun. Multi-label learning by exploiting label dependency [C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 999-1007.

[6] BOUTELL M R, LUO Jiebo, SHEN Xipeng, et al. Learning multi-label scene classification [J]. Pattern Recognition, 2004, 37(9): 1757-1771.

[7] FURNKRANZ J, HULLERMEIER E, MENCIA E L, et al. Multi-label classification via calibrated label ranking [J]. Machine Learning, 2008, 73(2): 133-153.

[8] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multilabel classification[C]//Proceedings of the 18th European Conference on Machine Learning. Berlin: Springer, 2007: 406-417.

[9] ZHANG Minling, ZHOU Zhihua. ML-kNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[10] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification[M]//DIETTERICH T G, BECKER S, GHAHRAMANI Z. Advances in Neural Information

Processing Systems 14. Cambridge, USA: The MIT Press, 2002: 681-687.

[11] LIU Yi, JIN Rong, YANG Liu. Semi-supervised multi-label learning by constrained non-negative matrix factorization[C]//Proceedings of the 21st National Conference on Artificial Intelligence. Menlo Park, USA, 2006: 421-426.

[12] 姜远,佘俏俏,黎铭,等.一种直推式多标记文档分类方法[J].计算机研究与发展, 2008, 45(11): 1817-1823.
JIANG Yuan, SHE Qiaoqiao, LI Ming, et al. A transductive multi-label text categorization approach[J]. Journal of Computer Research and Development, 2008, 45(11): 1817-1823.

[13] CHEN Gang, SONG Yangqiu, WANG Fei, et al. Semi-supervised multi-label learning by solving a Sylvester equation[C]//Proceedings of SIAM International Conference on Data Mining. Los Alamitos, USA, 2008: 410-419.

[14] SUN Yuyin, ZHANG Yin, ZHOU Zhihua. Multi-label learning with weak label[C]//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Menlo Park, USA, 2010: 593-598.

[15] 孔祥南,黎铭,姜远,等.一种针对弱标记的直推式多标记分类方法[J].计算机研究与发展, 2010, 47(8): 1392-1399.
KONG Xiangnan, LI Ming, JIANG Yuan, et al. A transductive multi-label classification method for weak labeling[J]. Journal of Computer Research and Development, 2010, 47(8): 1392-1399.

[16] KONG Xiangnan, NG M K, ZHOU Zhihua. Transductive multi-label learning via label set propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 704-719.

[17] 李宇峰,黄圣君,周志华.一种基于正则化的半监督多标记学习方法[J].计算机研究与发展, 2012, 49(6): 1272-1278.
LI Yufeng, HUANG Shengjun, ZHOU Zhihua. Regularized semi-supervised multi-label learning[J]. Journal of Computer Research and Development, 2012, 49(6): 1272-1278.

[18] 周志华,王珏.机器学习及其应用[M].北京:清华大学出版社, 2007: 259-275.

[19] ZHOU Zhihua, LI Ming. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541.

[20] Multi-label datasets[EB/OL]. [2013-01-06]. http://sourceforge.net/projects/mulan/files/datasets/.

作者简介:

刘杨磊,男,1990 年生,硕士研究生,主要研究方向为机器学习.发表学术论文 3 篇,获得计算机软件著作权登记 3 项.

梁吉业,男,1962 年生,教授,博士生导师,博士,主要研究方向为机器学习、计算智能、数据挖掘等.先后主持国家自然科学基金重点项目 1 项、国家“863”计划项目 2 项,国家“973”计划前期研究专项 1 项、国家自然科学基金项目 4 项.发表学术论文 150 余篇,出版著作 2 部,获发明专利 8 项.

高嘉伟,男,1980 年生,讲师,主要研究方向为机器学习.参与国家“863”计划项目 1 项、国家自然科学基金项目 3 项和山西省自然科学基金项目 4 项,发表学术论文 10 余篇.