一种增强差异性的半监督协同分类算法

于重重[1,2],商利利[2],谭 励[2],涂序彦[1],杨 扬[1],王竞燕[2]
(1.北京科技大学计算机与通信工程学院,北京 100083;2.北京工商大学计算机与信息工程学院,北京 100048)

摘 要: 半监督学习中的 Tri-Training 算法打破了以往算法对充分冗余视图的限制,并通过利用三个分类器处理标记置信度和样本预测问题提高了标记效率.为进一步增强协同训练过程中分类器之间的差异性以提高性能,本文在其理论基础上提出了一种增强差异性的半监督协同分类算法.该算法利用三个不同的分类器进行学习;考虑到分类模型在更新过程中,可能会因随机抽样导致性能恶化,该算法利用基于标记类别的分层抽样法来对已标记样本集进行抽样,并通过基于分类正确率的加权投票法实现了分类器的集成,提高了预测准确率.本文通过实验对所提出算法与 Tri-Training 算法做了性能比较,实验结果表明本文所提出的方法在分类问题上具有较好的性能,验证了该算法的有效性和可行性.

关键词: 半监督协同分类算法; Tri-Training 算法; 增强差异性策略; 分层抽样法
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2013)01-0035-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.01.007

A Semi-supervised Collaboration Classification Algorithm with Enhanced Difference

YU Chong-chong[1,2], SHANG Li-li[2], TAN Li[2], TU Xu-yan[1], YANG Yang[1], WANG Jing-yan[2]
(1. *School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*;
2. *School of Computer and Information Engineering, Beijing Technology and Business University, Beijing* 100048, *China*)

Abstract: Tri-Training algorithm in semi-supervised learning broke the restriction of previous algorithms on sufficient and redundant views and raised label efficiency by applying three classifiers to deal with labeled confidence. In order to further improve classifiers' performance through enhancing the difference between them, a semi-supervised collaborative classification algorithm with enhanced difference that applies three different classifiers was presented in this paper. Taking the might-be performance deterioration led by random sampling during the updating of classifying models into consideration, a method of stratified sampling based on labeled class was used by the algorithm to sample from the labeled sample sets, and the method of weighted voting based on classification accuracy realized the classifier ensemble, as a result the prediction accuracy is raised. Performance comparison between the proposed algorithm and Tri-Training algorithm was made through experiments, and the results show effectiveness of the former.

Key words: semi-supervised collaboration classification algorithm; Tri-Training algorithm; strategy of enhancing difference; stratified sampling

1 引言

在半监督学习算法中,Co-training 是一类很重要的算法[1].最初的 Co-training 算法是由 A Blum 和 T Mitchell[2]提出的,该算法适用于两个充分冗余视图的数据属性集,Nigam 和 Ghani[3]说明了 Co-training 算法中对两个假设条件要求的严格性.然而现实中所要解决的问题往往很难满足充分冗余视图这一前提条件,而且视图间的条件独立的要求也限制了 Co-training 算法的应用范围.在以后对于该类算法的研究中,不同的学者针对其中的不同问题进行研究.多数是在标准 Co-training 的基础上,对如下三个问题的解决:(1)标准 Co-training 算法中两个假设(视图充分冗余且相互独立)条件的难以满足;(2)训练过程中如何保持分类器之间的差异性;(3)标记时选择未标记样本的方法(置信度计算方法).

侯、焦[4]提出一种基于图的 Co-training 网页分类算

收稿日期:2012-08-06;修回日期:2012-10-21
基金项目:国家自然科学基金(No.61070182);北京市组织部优秀人才资助项目(No.2010D005003000008);北京市学科建设项目(No.PXM2012_014213_0000_74);北京市学科建设项目(No.pxm_2012_014213_000023)

法,在 Co-training 算法框架下,迭代地学习一个基于由超链接信息构造的图的半监督分类器和一个基于文本特征的 Bayes 分类器.Nigam 和 Ghani 提出的 CO-EM 算法[3]结合了 Co-training 算法和 EM 算法,通过天然特征分割得到两个视图.后来,Goldman 和 Y Zhou[5]提出了一种属性集不需要充分冗余视图的 Co-training 算法,称为 Statistical Co-Learning.该算法中使用了不同的决策树,并且由同一个数据属性集训练得到两个不同的分类器,这样使得每个分类器将样本空间划分为若干个等价类.为了进一步放松约束条件,Z H Zhou 和 M Li[6]提出了一种既不要求充分冗余视图、也不要求使用不同类型分类器的 Tri-Training 算法.由 Schwenker 等人[7]提出的半监督学习框架 Co-training by Committee(CoBC)通过多个差异性分类器的集成学习解决了 Co-training 算法中关于充分冗余和相互独立视图的不实际要求.唐等人[8]通过分析 Co- training 的理论假设,把寻找两个满足一致性和独立性特征视图的目标,转变成寻找两个既满足一定准确性,又存在较大差异两个的分类器,并根据每个视图上采用的分类算法是否相同,提出了两种改进算法 TV-SC 和 TV-DC.王、罗等人提出一种基于随机子空间的协同训练算法[9],其在数据特征空间的随机子空间中训练分类器,用一个分类器最置信的数据扩大其它分类器的训练集,以此提高分类器性能.

本文拟综合考虑上述三方面的问题,提出较为全面的解决方法.众所周知,分类器的多样性是提高分类性能的关键[10].在研究了 Tri-Training 算法的基础上,本文提出的增强差异性的半监督协同分类算法(A Semi-supervised Collaboration Classification Algorithm with Enhanced Difference,简称 DSCC 算法),通过对分类器的多样化、基于相似度的样本抽取等途径来增强多分类器之间的差异性,利用基于标记类别的分层抽样改进了模型更新条件,并通过基于分类正确率的加权投票法来提高对未标记样本预测的正确率.因此该算法的基本内容主要包括分类器的多样化、样本标记、分层抽样和加权预测四个部分.最后本文还通过实验验证了 DSCC 算法的有效性和可行性.

2 增强差异性的半监督协同分类算法

DSCC 算法首先利用三种不同的监督算法来产生初始分类器.在训练过程中,每个分类器训练集中的新标记样本都是由其余两个分类器协作提供.模型更新时,若不满足更新条件,则采用基于标记类别的分层抽样方法来对对已标记样本进行抽样,以样本容量小的子集代替已标记样本来重新判断,增加了进行下一次迭代的可能性.利用分类模型对未标记数据进行预测时,考虑到三个分类器之间的差异性,采用基于分类正确率的加权投票法来实现分类器的集成,从而提高预测结果准确率.

具体来说,给定未标记数据集 U,已标记数据集 L,DSCC 算法按如下步骤执行:

(1)初始化

对 L 进行自助抽样,产生的三个子集 S_1, S_2, S_3 分别作为训练集,利用不同的监督算法产生三种初始分类器 H_1, H_2, H_3.

(2)协同训练过程

第 t 次循环:

①对分类器 $H_i(i=1,2,3)$,利用另外两个分类器 $H_j, H_k(j,k=1,2,3$ 且 $j,k\neq i,)$对未标记数据进行标记,选择满足条件 $L_i=\{x \mid x\in U$ 且 $H_j(x)=H_k(x)\}$ 的样本生成新的标记样本 $S_i^{'}=L\cup L_i$;

②若本轮错误率小于上一轮且满足式(1)中所示条件,则利用 $S_i^{'}$ 更新分类器 H_i,否则利用基于标记类别的分层抽样方法对已标记样本集 L^t 进行抽样产生抽样子集 L_{sub}^t,以 L_{sub}^t 代替 L^t 重新判断条件,若满足转至①,否则迭代终止;

$$0<\frac{e^t}{e^{t-1}}<\frac{|L^{t-1}|}{|L^t|}<1 \tag{1}$$

式中 e^t 和 e^{t-1} 分别为分类器在第 t 轮和第 $t-1$ 轮迭代的分类错误率,$|L^t|$ 和 $|L^{t-1}|$ 分别为在第 t 轮和第 $t-1$ 轮迭代时分类器的已标记样本集大小.

第 t 次循环结束.

(3)未标记样本预测过程

对任意给定的未标记数据采用集成分类器 $\{H_1, H_2, H_3\}$ 按照基于分类准确率的加权投票法进行标记,根据权值分配方法计算三分类器的权值 w_1, w_2, w_3,三分类器对未标记样本进行标记,标记结果为 y_1, y_2, y_3,则未标记样本的最终标记 y 根据式(2)计算.

$$y=\text{sign}(w_1y_1+w_2y_2+w_3y_3) \tag{2}$$

式中,$\text{sign}(x)$ 为符号函数.

本文算法相对于 Tri-Training 算法而言,主要有以下优势:

(1)三个不同的分类器的使用,保证了训练过程中分类器多样性的保持;

(2)模型更新条件的改进,避免了因样本选取不当而造成分类性能恶化的问题;

(3)预测过程中采用基于分类准确率的加权投票法进一步提高了算法准确率和泛化能力.

下面从分类器多样化形式、模型更新策略和未标记样本预测方法三个方面阐述 DSCC 算法.

3 DSCC 算法的分类器多样化形式

Tri-Training 算法中选用三个相同的分类器,为了保

证分类器之间的差异性,对原始已标记样本进行自助抽样来产生三个已标记样本集.由此可能会产生以下两个问题:

(1)对不平衡数据集的适用性

现实问题中的数据样本往往存在样本不平衡问题.多数情况下,数据集中的少数类样本具有更高的"价值",需要更多"关注".这种样本集数据本身差异性很小,仅通过初始样本的采样并不能很好的实现分类器的多样性,反而使得多分类器的协同训练退化为单分类器的自训练,失去了半监督学习和协同训练的价值和意义.

(2)多分类器使用的有效性

使用同一种学习方法对数据进行训练,即使训练样本集不同,对于任意一个未标记样本,如果在实际中其正确标记是 p,那么无论用的是何种分类算法,只要在学习得到的分类器正确率较高的情况下,分类器对样本的分类结果相同的概率要大于不同的概率,因此会减弱使用多分类器的意义,同时也会导致算法的泛化能力不强.

针对上述问题.本文提出选用不同的监督分类算法作为学习算法产生三个基分类器,这不但可以解决因样本集差异性小而导致的多样性缺乏问题,提高了算法对不平衡数据集的适用性,而且还降低了分类器对任意两样本标记结果相同的概率.

4　DSCC 算法的模型更新策略

在使用三个分类器进行协同训练的多次迭代过程中,根据 Tri-Training 算法的思想[6],只要每个分类器的分类错误率还在减小,已标记样本集的容量还在扩大,那么分类器模型还将继续更新,协同训练的过程将不会停止.此外,在第 t 轮的更新中还需要满足式(1).

不等式条件不但需要分类器满足本轮错误率小于上一轮,即 $e^t < e^{t-1}$,而且需要满足 $e^t|L^t| < e^{t-1}|L^{t-1}|$.注意到 $e^t < e^{t-1}$ 且 $|L^t| > |L^{t-1}|$ 时,由于 $|L^t|$ 可能远远大于 $|L^{t-1}|$,因此 $e^t|L^t|$ 可能并不小于 $e^{t-1}|L^{t-1}|$.为了保证在这种情况下 $e^t|L^t| < e^{t-1}|L^{t-1}|$ 仍然成立,Tri-Training 算法中采用的是对 L^t 进行随机抽样的方法.令 s 代表采样后 L^t 的大小,那么若要满足 $e^t|L^t| < e^{t-1}|L^{t-1}|$,只需:

$$s = \left\lceil \frac{|L^{t-1}|}{|L^t|} - 1 \right\rceil \tag{3}$$

其中 L^{t-1} 应满足:

$$|L^{t-1}| > \frac{e^t}{e^{t-1} - e^t} \tag{4}$$

使得子采样后仍有 $|L^t| > |L^{t-1}|$.

由此产生了对于第 t 轮标记样本 L^t 的抽样问题.在这种状况下,Tri-Training 算法对于样本 L^t 采用随机抽样的做法,可能会因为样本选取不当而使模型在第 t 轮更新后,并不如 $t-1$ 轮的模型.尤其是当样本集本身具有不平衡特性时,已标记的样本中可能含有大量的相似样本,这类样本可能具有相同的标记(如均为正样本),或者样本属性数据具有相似的分布.对这类样本集进行随机子抽样,会造成因样本差异性小使得分类器准确率低,得到的分类模型的泛化能力不强.

针对上述问题,DSCC 算法采用基于标记类别的分层抽样方法对 L^t 进行抽样,以保证抽样子集中样本的多样性和差异性.

4.1　基于标记类别的分层抽样方法

考虑到分层抽样的优点,为了得到能够较为全面反映样本特征的样本子集,本文提出一种基于标记类别的分层抽样方法.具体做法是首先统计已标记样本中类别的数量,记为 H,H 即代表了分层抽样的层数.统计各种标记类别的样本比例,并按照比例和抽样子集的大小确定每一层的抽样数量.用伪代码描述该方法的基本过程如下面算法所示.

算法:基于标记类别的分层抽样算法

输入:L:已标记样本集

　　n_s:抽样子集的样本容量

输出:S:从 L 中选出的抽样子集 n_s

```
S←Φ
for 每一个样本 x∈L do   H = H∪H_x % 统计标记类别数量
end for
将 L 根据类别属性分成 H 层,第 i 层的样本数量为 L_i
for i = 1:H do
    n_i = L_i * n_s / H   % 确定第 i 层抽样数量 n_i
    S_i = Subsample(n_i, L_i)   % 抽样
    S = S_i ∪ S
end for
return S;
```

4.2　模型更新策略流程

在 DSCC 算法的执行过程中,三个分类器的更新是逐一进行的,下面给出其中一个在第 t 轮迭代中的更新流程:

(1)对标记样本集 L 进行标记,计算本次迭代的样本标记错误率 e^t;

(2)比较本轮与上一轮样本标记错误率 e^{t-1} 的大小,若 $e^t > e^{t-1}$,该分类器更新结束;否则转向(3);

(3)判断本轮与上一轮新标记的样本容量大小,若 $|L^t| \leqslant |L^{t-1}|$,该分类器更新结束;否则转向(4);

(4)若 $e^t|L^t| < e^{t-1}|L^{t-1}|$,将将新标记样本 L^t 与已标记样本 L 作为训练集,更新模型;否则转向(5);

(5)对 L^t 进行基于标记类别的分层抽样产生样本子集 L^t_{sub}使其满足 $e^t|L^t_{sub}|<e^{t-1}|L^{t-1}|$,将 L^t_{sub}与 L 一起作为训练集,更新模型.

5 DSCC 算法的未标记样本预测方法

DSCC 算法采用了三种不同的分类器,对测试样本进行标记时为保证标记结果的准确性,本文提出了基于分类正确率的加权投票法.其核心思想是对于给定样本集,分类正确率高的赋予较高的权值,分类正确率相对较低的赋予低的权值.其中分类正确率采用的是分类器对于已标记样本 L 的分类正确率.最终的预测结果利用式(1)来得到.

设有组合预测模型中共有 m 个预测模型,其预测误差分别为 $e_i(i=1,2,\cdots,m)$,排序后的序号分别为 num_i.要计算每个预测模型的权重,首先把单项预测模型的预测误差进行排序,然后基于上述事实,并依据简单加权平均和二项式系数的基本概念,得到三种权重计算方法,分别为:

①相对误差倒数法

$$w_i=\frac{e_i^{-1}}{e_1^{-1}+e_2^{-1}+\cdots+e_m^{-1}},i=1,2,\cdots,m \tag{5}$$

②加权平均方法

$$w_i=\frac{num_i}{\sum_{i-1}^{m}i}=\frac{2num_i}{m(m+1)},i=1,2,\cdots,m \tag{6}$$

③二项式系数法

$$w_i=\frac{C_{2m-1}^{i-1}}{2^{2m-1}},i=0,1,2,\cdots,m-1 \tag{7}$$

根据三种权值分配方法的基本内容,相对误差倒数法所获得的权值是连续值,而后两者所能分配的权值是离散值,第一种方法能更为精确地反映出分类器之间的性能差异.且相对误差倒数法中权值与预测误差成反比的赋权思想同前文中基于分类正确率的加权投票法的核心思想保持一致,因此从理论上来说第一种方法为最合适的权重分配方法.

综上所述,DSCC 算法中,采用相对误差倒数法进行分类器的权值分配公式如下:

$$w_i=\frac{e_i^{-1}}{e_1^{-1}+e_2^{-1}+e_3^{-1}},i=1,2,3 \tag{8}$$

其中 e_i 为分类器 H_i 的分类误差.

6 验证实验

前文在阐述 DSCC 算法时主要着眼于理论上的分析,下面进一步通过实验来验证本文提出算法的有效性和可行性.另外,经过分析对比我们发现理论上来说相对误差倒数法是比较合适的方法,在本实验中会进一步验证这个结论.

6.1 实验条件

(1)数据来源

本实验中的数据来源于 UCI 标准数据集.数据集的具体特征如表 1 所示.

表 1 数据集特征统计表

Dataset	attribute	size	class	pos/neg
australian	14	690	2	55.5%/44.5%
bupa	6	345	2	42.0%/58.0%
colic	22	368	2	63.0%/37.0%
diabetes	8	768	2	65.1%/34.9%
german	20	1000	2	70.0%/30.0%
hypothyroid	25	3163	2	4.8%/95.2%
ionosphere	34	351	2	35.9%/64.1%
kr-vs-kp	36	3196	2	52.2%/47.8%
sick	29	3772	2	6.1%/93.9%
tic-tac-toe	9	958	2	65.3%/34.7%
vote	16	435	2	61.4%/38.6%
wdbc	30	569	2	37.3%/62.7%

(2)样本分配

对于每一个数据集,选用 25% 的数据作为测试样本集,剩余的 75% 的数据作为训练样本集,其中,训练集中未标记样本的比例依次选用 20%,40%,60%,80%.

(3)分类器选择

DSCC 算法中的三个分类器分别选择 J48(C4.5)、RBF 和 NaiveBayes.比较算法为 Tri-Training 算法,由于该算法采用三个相同的分类器,为了进行全面的比较考虑了其分类器选用 J48(C4.5)、RBF 和 NaiveBayes 的三种情况.

(4)算法评价指标

采用算法的分类错误率作为算法评价指标,并定义性能提高比率做进一步的比较分析.

$$R=\frac{N_{correct}}{N}\times 100\% \tag{9}$$

$$R_{improve}=\frac{e_1-e_2}{e}\times 100\% \tag{10}$$

其中,R 指的是算法分类错误率,$N_{correct}$代表分类正确的样本数目,N 指的是总样本数;$R_{improve}$代表算法 2 相对于算法 1 的性能提高比率.

6.2 实验结果

所记录的实验结果中,在权值分配方法上分别采用相对误差倒数法、简单加权平均法和二项式系数法.

(1)算法错误率

表 2 ~ 表 5 分别记录了在不同未标记率下,基于三种权值分配方法 DSCC 算法和采用三种不同分类器的 Tri-Training 算法的分类错误率.

表 2　20% 未标记样本率

Dataset	Tri-Training			DSCC		
	J48	RBF	NB	RW	SW	BW
australian	.1462	.2047	.1871	.1404	.1404	.1287
bupa	.3140	.3953	.5116	.3488	.3372	.3372
colic	.1848	.2717	.2500	.1739	.1848	.1739
diabetes	.2396	.2500	.2188	.2240	.2292	.2240
german	.3240	.2440	.2280	.3160	.3080	.3120
hypothyroid	.1730	.2004	.1055	.1519	.1561	.1519
ionosphere	.0455	.0568	.3864	.1136	.1136	.1136
kr-vs-kp	.0271	.1412	.1323	.0527	.0416	.0571
sick	.0647	.0724	.0812	.0717	.0621	.0635
tic-tac-toe	.1131	.2527	.2701	.1138	.1257	.1258
vote	.0636	.100	.1364	.0727	.0545	.0727
wdbc	.0140	.0420	.0560	.0560	.0560	.0560

表 3　40% 未标记样本率

Dataset	Tri-Training			DSCC		
	J48	RBF	NB	RW	SW	BW
australian	.1754	.1813	.1871	.1754	.1930	.1871
bupa	.2907	.3837	.5233	.3721	.3721	.3721
colic	.1956	.2826	.2609	.2065	.2174	.2174
diabetes	.2604	.2708	.2344	.2240	.2604	.2500
german	.3360	.2400	.2680	.3280	.3000	.3120
hypothyroid	.0274	.0781	.0759	.0295	.0379	.0295
ionosphere	.1477	.1023	.3864	.1818	.1932	.1704
kr-vs-kp	.0283	.1517	.1415	.0654	.1284	.1369
sick	.1123	.1317	.1412	.1214	.1132	.1133
tic-tac-toe	.1274	.2712	.2987	.1321	.2512	.2437
vote	.1000	.0909	.1364	.1091	.1000	.1091
wdbc	.1888	.1399	.0560	.2098	.2238	.2098

表 4　60% 未标记样本率

Dataset	Tri-Training			DSCC		
	J48	RBF	NB	RW	SW	BW
australian	.1637	.1813	.2105	.1696	.1696	.1696
bupa	.3140	.5581	.5184	.3953	.4186	.3953
colic	.2282	.2391	.1957	.2391	.2174	.2391
diabetes	.2552	.2552	.2552	.2604	.2292	.2448
german	.3560	.2720	.2360	.3560	.3000	.3480
hypothyroid	.0190	.0696	.0654	.0211	.0189	.0189
ionosphere	.0227	.1591	.2727	.2841	.2727	.2727
kr-vs-kp	.0301	.1492	.1334	.0987	.1365	.1224
sick	.1241	.1050	.1336	.1326	.1230	.1283
tic-tac-toe	.1537	.2803	.3012	.1624	.1532	.1524
vote	.0727	.1455	.1364	.1364	.1273	.1364
wdbc	.2727	.0490	.0350	.1189	.1189	.1189

表 5　80% 未标记样本率

Dataset	Tri-Training			DSCC		
	J48	RBF	NB	RW	SW	BW
australian	.1579	.2222	.2339	.1404	.1345	.1345
bupa	.4125	.4563	.4721	.4213	.4017	.4068
colic	.1848	.3043	.2609	.2500	.2826	.2935
diabetes	.3385	.2917	.2708	.3385	.3490	.3281
german	.3320	.2920	.2560	.3460	.3280	.3480
hypothyroid	.0084	.0274	.0443	.0063	.0063	.0063
ionosphere	.0909	.0568	.3636	.2500	.2727	.2727
kr-vs-kp	.0316	.1523	.1407	.0854	.1470	.3650
sick	.1177	.1283	.1340	.1294	.1166	.1188
tic-tac-toe	.1419	.3137	.3317	.1650	.2987	.3124
vote	.0818	.1273	.1364	.1364	.1364	.1364
wdbc	.1189	.1469	.0629	.2098	.2238	.2098

从表 2 ~ 表 5 的结果可以得到如下结论:

①对于不同的样本集,基于三种权值分配方法的 DSCC 算法的分类性能没有绝对的优劣之分.原因是权重是根据分类器对于标记样本 L 的分类错误率来分配的,当 L 占的比率很小时,不足以反映出分类器之间对于样本的分类性能差距.

②通过数据集 hypothyroid 和 sick 这两种不平衡数据集的实验结果可以看出,本文算法针对不平衡数据集分类具有较好的效果.

③根据算法对同一数据集的分类误差率变化趋势可以看出,基于相对误差倒数法的 DSCC 算法的分类性能表现最为稳定,且误差率相对较小,这与前文的理论分析保持一致.因此,权值分配方法采用误差平方和倒数方法.

(2)算法的相对性能提高比率

算法 DSCC 相对于 Tri-Training 算法的性能提高比率为:

$$R_{\text{improve}} = \frac{e_{\text{TT}} - e_{\text{DSCC}}}{e_{\text{TT}}} \times 100\% \tag{11}$$

根据表 2 ~ 表 5 的算法错误率,计算基于相对误差倒数法的 DSCC 算法相对于 Tri-Training 算法的性能提高比率,比较结果如图 1 示(图中 TT-J48,TT-RBF 和 TT-NB 分别代表采用 J48、RBF 和 NaiveBayes 分类器的 Tri-Training 算法).

由图 1 可以得到如下结论:

①在大多数情况下,增强差异性的半监督协同分类算法相对 Tri-Training 算法的性能有一定的提高.

②增强差异性的半监督协同分类算法相对于 Tri-Training 算法而言,在不同的未标记样本率下,性能提高比率变化曲线呈现出了一致的变化趋势,这说明了 DSCC 算法的性能稳定性及有效性.

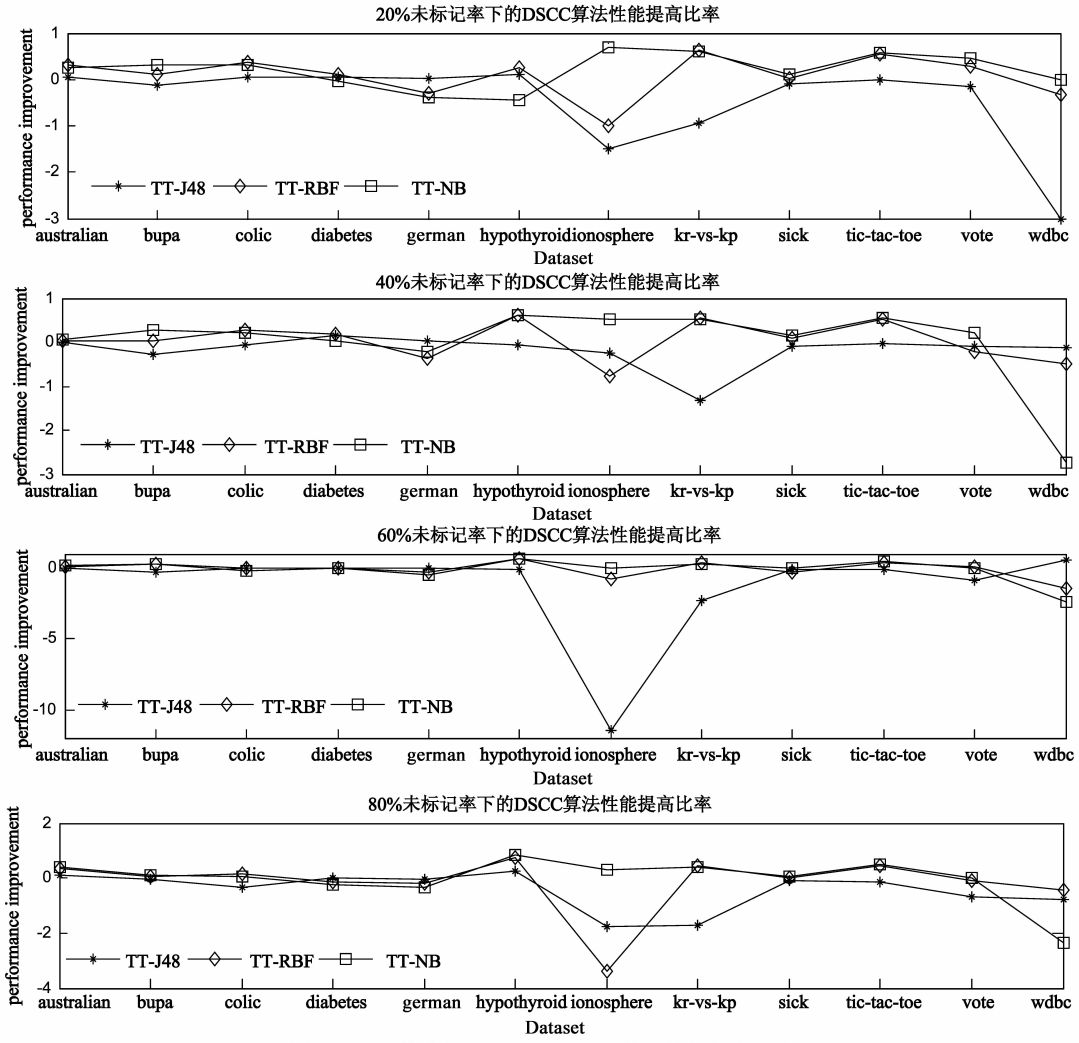

图1 DSCC算法相对于Tri-Training算法性能提高比率

7 总结与展望

在半监督学习中,Co-training 算法成功的关键在于保持分类器之间的多样性,DSCC 算法采用三种不同的分类器使其在训练过程中很好的保持多样性,更充分的获取未标记样本中的信息.该算法在模型更新时为抽取更合适的样本子集,最优化分类器模型的性能,采用了基于标记类别的分层样本抽样法.此外,对未标记样本的预测则是通过基于分类正确率的加权投票法来实现,且通过实验确定了分类器权重的分配方法.为了验证 DSCC 算法的有效性及合理性,本文除了理论分析之外,还进行了详细的实验对比.实验结果验证了本文算法的适用性和有效性.

本文给出了 DSCC 算法的具体内容及实现方式,其中包括理论支持和实验验证.希望在以后的具体应用中进一步探索 DSCC 算法在不同领域的性能特点.

参考文献

[1] 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.
Zhi-Hua Zhou, Yu Wang. Machine Learning and Its Applications[M]. Beijing: Tsinghua University Press, 2007.259 – 275. (in Chinese)

[2] A Blum, T Mitchell. Combining labeled and unlabeled data with co-training[A]. Proceedings of the Eleventh Annual Conference on Computational Learning Theory [C]. Berlin, German: Springer, 1998.92 – 100.

[3] K Nigam, R Ghani. Analyzing the effectiveness and applicability of co-training[A]. Proceedings of the 9th International Conference on Information and Knowledge Management [C]. New York, USA: ACM, 2000.86 – 93.

[4] 侯翠琴,焦李成.基于图的 co-training 网页分类[J].电子学报,2009,37(10):2173 – 2219.
HOU Cui-qin, JIAO Li-cheng. Graph based co-training algo-

rithm for web page classification[J]. Acta Electronica Sinica, 2009,37(10):2173 – 2219.(in Chinese)

[5] S Goldman, Y Zhou. Enhancing supervised learning with unlabeled data[A]. Proceedings of the 17th International Conference on Machine Learning[C]. San Francisco: Morgan Kaufmann,2000.327 – 334.

[6] ZHOU Z H, LI M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(11):1529 – 1541.

[7] M F Abdel Hady, F Schwenker. Co-training by committee: A new semi-supervised learning framework[A]. Proceedings of the IEEE International Conference on Data Mining Workshops [C]. Piscataway : IEEE,2008.563 – 572.

[8] 唐焕玲,林正奎,鲁明羽.基于差异性评估对 Co-training 文本分类算法的改进[J].电子学报,2009,36(12A):138 – 143.
TANG Huan-ling, LIN Zheng-kui, LU Ming-yu. An improved co-training text categorization algorithm based on diversity measures[J]. Acta Electronica Sinica, 2009, 36(12A): 138 – 143.(in Chinese)

[9] 王娇,罗四维,曾宪华.基于随机子空间的半监督协同训练算法[J].电子学报,2008,36(12A):60 – 65.
WANG Jiao, LUO Si-wei, ZENG Xian-hua. A random subspace method for co-training[J]. Acta Electronica Sinica,2008, 36(12A):60 – 65.(in Chinese)

[10] P Melville, R Mooney. Creating diversity in ensembles using artificial data[J]. Information Fusion,2004,6(1):99 – 111.

作者简介

于重重　女,1971 年 8 月生于重庆,北京工商大学计算机与信息工程学院教授、副院长、硕导,主要研究领域:智能信息处理与模式识别、复杂实时监测系统预测与评估.
E-mail: chongzhy@vip.sina.com

商利利　女,1986 年 12 月生于河南濮阳,北京工商大学计算机与信息工程学院在读硕士研究生,主要研究领域:机器学习、模式识别.
E-mail: shanglili2008v@126.com