基于信息反馈的半监督支持向量机算法①

杜红乐, 张 燕, 李 楠

(商洛学院 数学与计算机应用学院, 商洛 726000)

摘 要: 针对直推式支持向量机中标记速度与标注精度之间的矛盾, 提出一种信息反馈的半监督支持向量机算法, 该算法利用上轮标注数量、重置次数、未标注边界样本数量等信息, 动态调整标记样本数量, 对区域标注和成对标注进行折衷, 在继承渐进赋值和动态调整的同时, 可以平衡标记速度与标记精度之间的矛盾, 减少错误的传递和积累. 在人工数据集和 UCI 数据集上的实验结果表明该算法在保证标注准确度的前提下提高算法速度.

关键词: 支持向量机; 直推式学习; 半监督学习; 信息反馈

Semi-Supervised Support Vector Machine Algorithm Based on Information Feedback

DU Hong-Le, ZHANG Yan, LI Nan

(School of Mathematics and Computer Application, Shangluo University, Shangluo 726000, China)

Abstract: In order to resolve the contradiction between the speed and the precision of transductive support vector machine, a semi-supervised vector machine algorithm based on information feedback is proposed. The algorithm uses the information of the number of last round, the number of reset, the number of unlabeled samples to adjust dynamically the number of labeled samples, and make a tradeoff between region labeling and pairwise tagging. While the progressive evaluating and dynamically adjusting, it can balance the contradiction between the marking speed and accuracy and reduces the transmission and accumulation of errors. The experimental results on AI data sets and UCI data sets show that the proposed algorithm can improve calculation speed on the premise of ensuring the accuracy of label precision.

Key words: support vector machine; transductive learning; semi-supervised learning; information feedback

传统支持向量机是针对有监督学习, 即训练数据集是有标签样本, 而实际中由于样本标记比较困难或者标记费用较高导致有训练集中包含少量的有标签样本和大量的无标签样, 半监督支持向量机能够挖掘有标签样本包含的空间信息的同时, 还可以把无标签样本中的空间分布信息转移到最终的分类器中, 能大幅降低学习成本, 提高泛化性能, 因此半监督学习在机器学习领域备受关注. 半监督支持向量机学习算法主要有直推式支持向量机(TSVM)[1]、Ting 等提出的 BoostSVM 算法、Belkin 等提出的 LapSVM 算法. 由于 TSVM 自身的优点, 受到许多专家学者的关注[2-8], 也在很多领域得到应用[9-12].

文献[2]针对 TSVM 需要估计各类样本数目的问题进行改进, 基本思想是在 TSVM 训练过程中, 每次迭代只选取少量“重要”的为标记样本进行暂时标记, 在迭代过程中不断修正 TSVM, 该方法很好的实现了对有标签样本和无标签样本的学习, 但是存在训练速度慢、回溯次数过多等问题, 文献[3-8]提出了相应的改进措施, 文献[3-4]主要集中在对未标注样本的选取方法上, 文献[3]选择距离支持向量均值较远的样本进行标注, 同时采用动态调整半标定样本的惩罚参数; 文献[4]引入游程的概念, 利用游程标注那些可信度高的样本; 文献[5]主要对标记重置机制的改进, 对每个半标定样本增加双模糊度, 促进不一致样本的出现; 文献[6]在迭代时才用增量学习方法, 减少计算量, 提高速度.

以上方法主要针对标记准确度或者标记速度方面提高算法速度, 但在 TSVM 中, 标记速度与标记准确度之间是矛盾的, 成对标记法是每次选择标记准确可能性最大的两个样本进行标记, 使得标记的准确度较

① 基金项目:陕西省自然科学基础研究计划资助项目(2015JM6347);陕西省教育厅科技计划(15JK1218);商洛学院科学与技术研究项目(15sky010)

收稿时间:2016-09-07;收到修改稿时间:2016-10-16 [doi:10.15888/j.cnki.csa.005777]

高, 样本重置次数较少, 算法速度有保障; 区域标记法则按照标记准确度对样本进行排序, 选取若干个有高准确度的样本进行标记, 减少迭代次数, 保证算法速度. 每次迭代标记样本数少, 则迭代次数就会增加; 每次迭代样本数量多, 标记准确度就会降低导致样本标记错误率的增加, 一方面造成错误传递, 另一方面重置次数增加, 继而增加迭代次数. 随着数据收集技术的完善, 数据规模越来越大, 无标记样本数量也随之急增, 算法既需要考虑标记的准确度, 又需要考虑标记的速度.

基于以上分析, 针对半监督支持向量机算法在标注速度与标注准确度之间的矛盾, 本文提出一种基于信息反馈的半监督支持向量机算法(Information Feedback Semi-supervised Support Vector Machine, IFS3VM), 该算法依据已标注样本的标注是否准确决定标注速度, 使得标注速度与标注准确度之间达到一个平衡, 减少前期标注错误导致的错误传递, 在保证准确度的情况下提高算法的速度, 最后在人工数据集和 UCI 数据集上的实验结果进一步表明该算法的有效性.

1 直推式支持向量机

1.1 算法描述

直推式学习是对含有少量有标签样本和大量的无标签样本进行学习, 能够把无标签样本的空间分布信息转移到最终的分类器中, 由于无标签样本容易收集且数量较多, 包含更多的样本空间信息, 使得分类器具有较好的泛化能力. 因此直推式学习是许多专家学者关注的热点, 并在许多领域得到相应的应用. 目前最主要的研究成果是 Joachims 的直推式学习支持向量机(Transductive Support Vector Machine, TSVM), 随后有很多专家学者对 TSVM 进行相应的改进, 下面对 TSVM 算法实现进行简单介绍[1], 训练过程可以分为以下几个主要步骤:

步骤 1. 给定参数 C 和 C^*, 对有标签样本进行归纳学习, 得到最初分类器, 通过一定规则给定无标签样本的正标签样本数目 N.

步骤 2. 用初始分类器对无标签样本进行测试, 依据判别函数值, 对值最大的 N 个无标签样本暂时标记为正类, 其余的无标签样本标记为负类, 并给定临时惩罚因子 C^*_{temp}.

步骤 3. 把标记后的样本加入训练集中重新训练, 得到新的分类器, 然后按规则交换一对标记类别不同的测试样本的标记类别, 使得优化问题的目标函数值得到最大下降, 反复执行该步骤, 直到找不出符合交换条件的样本.

步骤 4. 均匀地增加临时惩罚因子的值, 返回到步骤 3, 当时算法结束.

1.2 样本标注方法

每次选择的标注样本对最终的结果有较大的影响, PTSVM 中采用成对标注法, 每次选择边界区域内的最大值和最小值进行标注, 成对标注法速度较慢, 提出区域标注法, 文献[6-9]都是对区域标注法的改进. 本文结合计算机网络中的慢启动拥塞避免算法思想, 采用一种依据标记结果进行信息反馈的自适应的样本标记方法. 每次训练完成后, 选择满足式(2)条件中的 K 个值最大的样本标注为正类, 选择满足式(3)条件的 K 个值最小的样本标注为负类.

$$\arg\max(f(x_i^*)), s.t. 0 < f(x_i^*) < 1 \quad (2)$$

$$\arg\min(f(x_i^*)), s.t. -1 < f(x_i^*) < 0 \quad (3)$$

其中 K 值决定了标注速度, 即直推式学习的学习速度, K＝1 即为成对标注法, K 取最大值时表示一次对边界区域中的样本全部标注, K 值可以通过寻优算法获得最优值, 本文针对标记后样本的分类情况进行信息反馈, 然后依据反馈信息确定 K 的取值. 介绍完标注样本选择方法后, 下面介绍每次迭代中分类器的构建.

2 信息反馈算法

2.1 反馈思想

TSVM 中标记速度和标记精度是一对矛盾, 成对标记法是选择未标记的边界样本值最大和最小的两个样本进行标记, 即选择标记准确率最大的两个样本, 准确率较高, 但是标记速度即训练速度较慢; 区域标记法是选择一定区域的样本进行标记, 每次标记的样本数量较多, 标记速度较快, 但是标记准确率较低.

在 TSVM 迭代过程中, 如果样本标记错误, 会对后续迭代中构建的分类超平面有一定的影响, 同时这个错误会被积累并传递下去, 为此在 TSVM 中, 若样本标记类别与之前标记类别不同, 则样本被重置(重新作为未标记样本), 换句话说, 样本标记类别可能不正确的情况下, 取消对该样本的标记, 减少标记错误及错误的传递. 然而, 成对标记法为了保证每次标记样

本类别的准确性，每次只选择最有把握的两个样本进行标记，为了提高标记速度,区域标记法每次迭代选择较多的样本进行标记，这样提高了算法速度，但会导致标记样本的准确性难于保证.

在网络传输中，传输速率和网络拥塞也是一对矛盾，传输速率是越大越好，但是传输速率越大，出现网络拥塞的可能性就越大，网络拥塞导致数据重传，数据重传会导致拥塞加剧，拥塞加剧导致重传次数增加，进一步导致有效传输速率的下降，拥塞严重时即网络瘫痪. 为此，在网络传输中采用拥塞控制，如慢启动、拥塞避免算法中，通过反馈拥塞信息进行动态调整传输速率. 网络传输速率与 TSVM 中标记速度相似，发生拥塞时对数据进行重传与 TSVM 中的样本标记错误时样本重置相似，因此依据网络拥塞控制思想，本文给出一种基于信息反馈的 TSVM 算法.

2.2 标注样本上限

样本被重置表明样本标记的类别可能不正确，造成标记错误的原因有: 1)分类器自身的原因，如学习不充分等; 2)每轮中标记样本过多(速度过大)，对标记可信度不大的样本也进行了标记. 对于第一类研究较多，这里不做赘述，对于第二类，廖东平等[4]把游程的概念引入到 TSVM 中，定义标记可信度，每次迭代中标记可信度大的样本，减少重置; 薛贞霞等[5]为提高样本标记的把握，每次选择距离各类支持向量平均距离最近的 1-2 个样本. 但是上面方法都是考虑标记速度并追求准确率，而没有考虑标记速度与标记准确性之间的冲突，本文将利用标记准确性来动态调整标记速度.

每轮循环能够标注的样本数影响算法的速度，标注样本数过多会导致标记速度过快，影响最终分类器的分类性能; 标注样本数过小会导致标记速度慢，值为 2 的时候就回到成对标注法. 本文依据网络中的慢启动拥塞避免算法思想，依据标记准确度动态调整样本标注数量，起始时选择 2 个样本进行标注，然后就是指数增加，当达到上限值(本文采用上限值为样本数的 10%)就不再增加，当有样本被重置，则标注样本数重新回到 2，重复上面过程直到算法终止.

2.3 算法描述

该算法结合网络拥塞避免算法思想和直推式支持向量机中标记速度与标记准确率冲突的问题，提出一种基于标记信息反馈的半监督支持向量机算法，若有样本需要重置，说明标记准确率存在问题，应该放慢标记速度，若没有样本需要重置，说明标记准确率较好，应适当提高标记速度，基于这种思想，提出基于信息反馈的半监督支持向量机算法，具体的算法的流程如图 1 所示.

结合图 1 的流程图，算法的具体过程描述如下:

算法: 信息反馈的半监督支持向量机算法 IFS3VM

输入: 有标签样本集 T，无标签样本集 U，分类器数目 m

输出: 最终分类器

Step1.对有标签样本集 T 进行训练，得到初始分类器 C_0;

Step2.用分类器分别对无标签样本集进行测试，输出每个样本的测试结果 $f_1, f_2, \cdots, f_l$;

Step3.判断前期样本标注类别与当前测试类别结果是否一致，如果标注一致，且 $2*k > k_{\max}$ 则 $k = 2*k$，若 $2*k \le k_{\max}$ 则 $k = k_{\max}$; 如果有样本标注不一致，则 $k = 1$，转 Step5;Step4.如果没有满足 $|f|_i < 1$ 的样本，则转 Step 8，否则进入 Step 5;

Step5.若满足 $|f|_i < 1$ 的样本数大于 K，则选择满足 $|f|_i < 1$ 的最大的 K 个样本进行标注，否则选取满足条件的全部样本，若 $f_i > 0$ 标注为正类; 若 $f_i < 0$ 标注为负类，把新标注的样本加入到训练集中，获得新训练集，转 Step6;

Step6.从训练集中删除标注不一致的样本，获得新训练集，转 Step7;

Step7.重新对训练集进行训练，得到分类器 C_i，返回 Step2;

Step8.训练得到最终分类器.

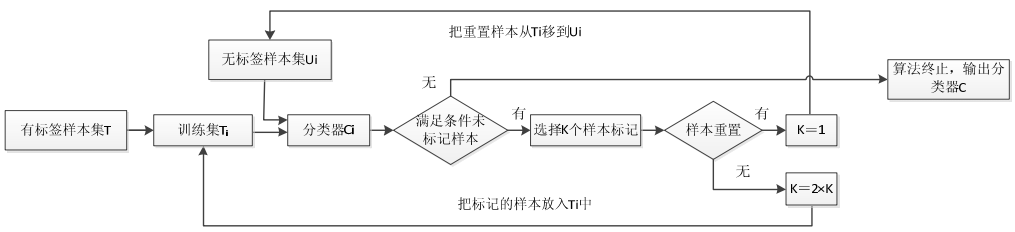

图 1　算法流程

3 实验及数据分析

本文算法是在 Matlab 7.11.0 环境下, 结合台湾林智仁老师的 LIBSVM[13]完成, 实验在 Intel Core i7 2.3GHz、4G 内存、操作系统为 Win7 的 PC 机上完成. 为验证算法的性能, 本文分别在人工数据集和 UCI 数据集上分别进行测试.

3.1 人工数据集

为便于观看数据集的变化, 该节采用人工数据集进行实验, 人工数据集分为线性可分和线性不可分两个数据集. 实验结果中 PTSVM 表示成对标记法、RLTSVM 表示区域标注法、IFS3VM 是本文的信息反馈法, 本文实验中的区域标注法是对两类边界样本各取一半进行标记(若边界样本数小于 10, 则全部标记).

3.1.1 线性可分

线性可分数据集为随机产生两类均匀分布的样本, 包括有标签样本 T 和无标签样本 U. 样本集 T 中, 第一类样本为 $U([0,1]\times[0,1])$, 第二类样本为 $U([0,1]\times[1,2])$, 两类样本各 10 个. 样本集 U 作为无标签样本, 同时作为测试样本, 第一类样本为 $U([0,1]\times[0,1])$, 第二类样本为 $U([0,1]\times[1,2])$, 两类样本各 100 个样本.

由于数据集具有随机性, 因此该节实验给出多次实验结果及平均值, 表 1 中给出了随机的 5 次实验结果及平均值, 其中一次准确率低于成对标记法外, 其它与成对标记法准确率相同, 实际中做了 20 次实验, 其中 3 次略低于成对标记法外, 其它与成对标记法相同. 观察样本分布图可以发现, 有标记样本的分布对训练时间和最终的准确度都有较大的影响.

表 1 实验结果对比表(1)

序号	PTSVM		RLTSVM		IFS3VM	
	训练时间(S)	准确率(%)	训练时间(S)	准确率(%)	训练时间(S)	准确率(%)
1	1.6297	95.75	0.1160	95	1.0410	95.75
2	1.6380	97.2	0.1310	96.25	0.9520	97
3	2.8860	98.8	0.2470	97	1.7931	98.8
4	9.8120	99.2	0.1810	98.8	2.3762	99.2
5	5.2260	98.8	0.0650	98.75	2.0850	98.8
平均值	4.2383	97.95	0.148	97.16	1.6495	97.91

为直观的观察样本被标记的过程, 图 2 给出了其中一次实验中分类超平面不断变化的图, 图中四幅图像分别是第 1 次、第 4 次、第 8 次、最后一次的分类超平面示意图. 由图中可以看出, 距离分类超平面近的样本比较稀疏, 这是因为算法在每次迭代过程中, 都是选择标记准确度最高及远离分类超平面的样本进行标记, 这样的优点就是每次标记的样本准确性较高, 但对分类超平面的影响相对也较小, 可能丢掉很重要的样本, 导致实际分类超平面与理想分类超平面有所偏移, 即分类超平面性能下降.

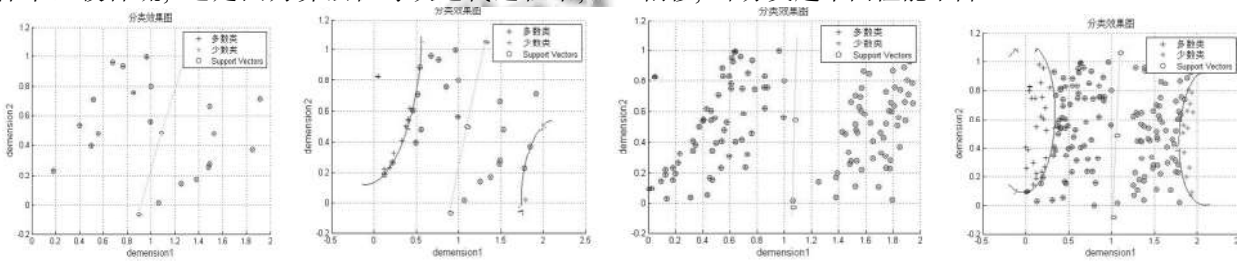

图 2 分类超平面

3.1.2 线性不可分

该部分实验数据产生两类同心圆样本 $\begin{cases}x=\rho g\cos\theta\\ y=\rho g\sin\theta\end{cases},\theta\in U[0,2\pi]$, 有标签样本的第一类样本的半径是均匀分布 $\rho\in[0.95,2]$, 第二类样本的半径是均匀分布 $\rho\in[0,1.05]$, 两类样本各 10 个; 无标签样本集和测试集为同一数据集, 与有标签样本具有相同的分布, 两类样本各 300 个. 这里选择径向基核函数 $K(x,y)=\exp(-g\|x-y\|^2)$.

由于数据集具有随机性, 因此该节实验给出多次实验结果及平均值, 表 2 中给出了随机的 5 次实验结果及平均值, 其中一次准确率低于成对标记法外, 其它与成对标记法准确率相同, 实际中做了 20 次实验, 其中 3 次略低于成对标记法外, 其它与成对标记法相同, 训练时间上远远小于成对标记算法. 观察样本分布图可以发现, 有标记样本的分布对训练时间和最终的准确度都有较大的影响.

表 2　实验结果对比表(2)

序号	PTSVM		RLTSVM		IFS3VM	
	训练时间(S)	准确率(%)	训练时间(S)	准确率(%)	训练时间(S)	准确率(%)
1	16.3721	94.6667	8.9310	92.3333	11.5470	94.6667
2	31.5830	92.8667	9.5270	91.8667	17.7580	92.8333
3	26.7940	90.3333	8.5960	90.3333	18.2920	90.3333
4	13.2580	91.3667	4.8710	90.8333	5.4530	91.3667
5	71.6310	91.3667	18.4690	90.3333	34.1560	91.3667
平均值	31.9276	92.02	10.0788	91.14	17.4412	92.1133

3.2 UCI 实验及结果分析

本实验数据集采用 Contraceptive Method Choice(Cmc)、haberman's survival、Ionosphere、Pima Indians Diabetes、balance 和 spice 六组 UCI 数据, 数据的分布情况如表 3 所示. 对选择的数据中随机选取 5% 的样本作为有标签样本, 用剩余 95%的样本作为无标签样本, 起始选取 2 个样本进行标记, 每轮循环中选择样本的上限为无标签样本总数的 10%(去整数部分), 无标签样本最后作为测试集.

为验证本文算法的有效性, 该小节分别对 PTSVM(成对标记直推式)、RLTSVM(区域标记直推式)、SVM(采用对原始数据集用支持向量机训练, 并对该数据集进行测试)及本文算法 IFS3VM 算法用表 3 中的数据进行实验, 分别从迭代次数、算法时间、准确率比较四种算法的性能.

表 3　实验数据集

序号	数据集	属性	多数类数量	类别数
1	Cmc	9	1473	3
2	haberman	3	306	2
3	ionosphere	34	351	2
4	pima	8	768	2
5	balance	4	625	2
6	spice	60	2175	2

表 4　准确率和训练时间对比

数据集	PTSVM		RLTSVM		IFS3VM		SVM	
	训练时间	准确率	训练时间	准确率	训练时间	准确率	训练时间	准确率
Cmc	4.1420	90.7076	1.1370	85.6326	1.9350	90.7076	0.1560	91.5818
haberman	0.1385	95.1034	0.0145	91.6263	0.8720	95.1034	0.0120	98.0392
ionosphere	1.9470	95.7421	0.5481	91.0811	1.0376	95.7421	0.0050	98.2906
pima	5.8190	97.3347	1.3910	94.3759	2.0850	97.3347	0.0930	100
balance	3.5840	95.7631	0.4382	93.3614	1.1480	95.7631	0.0160	98.0476
spice	6.6390	97.8964	2.8960	91.7348	2.8960	96.1868	0.5610	98.9540
平均值	2.7116	95.4246	1.0708	91.3020	1.6623	95.1396	0.1405	97.4855

由于有标签样本是随机选择的, 因此这里为了观察迭代次数采用 10 次实验的平均结果, 如表 4 所示. 结合表 4、表 5 可以看到, 算法对实验速度有了较大的提高, 实验准确率也较好, 该算法尤其对于数据规模较大时, 效果较明显.

表 5　迭代次数对比

数据集	PTSVM	RLTSVM	IFS3VM
Cmc	18.6	4.4	9.7
haberman	31.4	6.3	12.5
ionosphere	14.1	4.3	6.6

pima	11.8	4.9	4.7
balance	64.5	12.8	46.5
spice	75.1	13.7	57.6

4 结论

针对直推式支持向量机训练速度与标注准确度之间的冲突问题, 提出一种信息反馈的半监督支持向量机算法, 该方法依据网络拥塞控制中的慢启动拥塞避免算法的思想, 可以提高每次标注的准确率, 减少样本重置, 提高训练速度, 同时每次迭代训练集都是多个可以进一步提高训练速度, 实验结果也表明该算法的有效性. 如何在标记样本时即考虑标记准确性又考虑样本对分类超平面的贡献度, 提高最终分类器的分类性能是下阶段的主要工作.

参考文献

1 Vapnik VN. Statistical Learning Theory. New York: John Wiley and Sons, 1998.

2 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习算法.软件学报,2003,14(3):451–460.

3 王安娜,李云路,赵锋云,等.一种新的半监督直推式支持向量机分类算法.仪器仪表学报,2011,32(7):1546–1550.

4 汪海燕,黎建辉,杨风雷.支持向量机理论及算法研究综述.计算机应用研究,2014,31(5):1281–1286.

5 齐芳,冯昕,徐其江.基于人工鱼群优化的直推式支持向量机分类算法.计算机应用与软件,2013,30(3):294–296.

6 杜红乐.基于核空间中 K-近邻的不均衡数据算法.计算机科学与探索,2015,9(7):869–876.

7 王立梅,李金凤,岳琪.基于 k 均值聚类的直推式支持向量机学习算法.计算机工程与应用,2013,49(14):144–146.

8 齐芳,冯昕,徐其江.基于人工鱼群优化的直推式支持向量机分类算法.计算机应用与软件,2013,30(3):294–296.

9 杜红乐,张燕.密度不均衡数据分类算法, 西华大学学报(自然科学版),2015,34(5):16–23,74.

10 Bruzzone L, Mingmin CH, Marconcini M. A novel transductive SVM for semisupervised classification of remote-sensing images. IEEE Trans. on geoscience and remote sensing, 2006, 44(11): 3363–3373.

11 艾解清,高济,彭艳斌等.基于直推式支持向量机的协商决策模型.浙江大学学报(工学版),2012,46(6):967–973,994.

12 张建明,孙春梅,闫婷.基于自适应 SVM 的半监督主动学习视频标注.计算机工程,2013,39(8):190–195.

13 Chang CC, Lin CJ. LIBSVM: A library for support vector machines, 2015. Software available at http://www.csie.ntu.tw/~cjlin/libsvm.