代价敏感的半监督 Laplacian 支持向量机

万建武[1,2],杨 明[1,2],陈银娟[1]
(1. 南京师范大学计算机科学与技术学院,江苏南京 210046;
2. 南京师范大学数学科学学院,江苏南京 210046)

摘 要: 代价敏感学习是机器学习领域的一个研究热点.在实际应用中,数据集往往是不平衡的,存在着大量的无标签样本,只有少量的有标签样本,并且存在噪声.虽然针对该情况的代价敏感学习方法的研究已取得了一定的进展,但还需要进一步的深入研究.为此,本文提出了一种基于代价敏感的半监督 Laplacian 支持向量机.该模型在采用无标签扩展策略的基础上,将考虑了数据不平衡的错分代价融入到 Laplacian 支持向量机的经验损失和 Laplacian 正则化项中.考虑到噪声样本对决策平面的影响,本文定义了一种样本依赖的代价,对噪声样本赋予较低的权重.在 7 个 UCI 数据集和 8 个 NASA 软件数据集上的实验结果表明了本文算法的有效性.

关键词: 代价敏感学习; 半监督学习; Laplacian 支持向量机
中图分类号: TP39 **文献标识码:** A **文章编号:** 0372-2112 (2012) 07-1410-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.07.020

Cost Sensitive Semi-Supervised Laplacian Support Vector Machine

WAN Jian-wu[1,2], YANG Ming[1,2], CHEN Yin-juan[1]
(1. *School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu* 210046, *China*;
2. *School of Mathematics Science, Nanjing Normal University, Nanjing, Jiangsu* 210046, *China*)

Abstract: Cost sensitive learning is the hot research area in machine learning. In practical real applications, the datasets are usually class-imbalanced, most of the samples are unlabeled, only a few of the samples are labeled, and noise samples are existed. Although some progress has been made in cost sensitive learning for such situation, it needs further solved. For that we propose a semi-supervised Laplacian support vector machine based on cost sensitive learning. On the basis of label propagation, the proposed model integrates the misclassification costs considering class-imbalance into the hinge loss and Laplacian regularization of the Laplacian support vector machine. At the same time, considering the effect on the decision hypersphere of the noise samples, we define an example-dependent cost which makes the weights of noise samples lower. The experimental results on 7 UCI, 8 NASA datasets demonstrate the superiority of our proposed algorithm.

Key words: cost sensitive learning; semi-supervised learning; Laplacian support vector machine

1 引言

代价敏感学习[1~10]是机器学习领域的一个研究热点.目前对代价敏感学习问题的研究主要集中于:研究数据不平衡性和代价之间的关系[1,3~5];推广现有的分类器模型,设计一个代价敏感的分类器[1,2,5,7];研究多类的代价敏感学习[9,10].

在代价敏感学习中,核心问题是代价.代价的种类有很多,例如:测试代价,学习代价,干预代价以及错分代价[10].目前研究最多的是错分代价.错分代价包括类依赖的代价和样本依赖的代价.类依赖的代价更容易给出,因此很多学者采用类依赖的代价[4~6,8,10].

在真实的数据集中,采集得到的数据往往是不平衡的,存在大量的无标签样本,只有少量的有标签样本,并且存在噪声.针对该情况,已有学者提出了一些解决方法.文献[4,5]的作者研究了数据不平衡性和代价之间的关系,提出了一个将代价敏感和数据不平衡统一学习的经验性框架,并在 UCI 数据集[11]上进行了验证.然而,现在有许多不平衡的数据集,例如 NASA 软件数据集[12],它是从美国宇航局用于卫星飞行控制以及地面控制的软件中提取出来的真实数据集.因此有必要在其他真实数据集(NASA 软件数据集)上对文献[5]提出的

收稿日期:2011-05-26;修回日期:2011-12-23
基金项目:国家自然科学基金(No.60873176);江苏省自然科学基金重点重大专项(No.BK2011005);江苏省自然科学基金(No.BK2011782)

模型进行验证.文献[6]的作者研究了半监督[8,13~16]情形下的代价敏感学习问题.文献[6]通过扩展无标签样本的标签信息,提高了算法的分类性能,但它没有考虑数据的不平衡性以及噪声问题.

为此,本文提出了一种基于代价敏感的半监督 Laplacian 支持向量机.该模型在采用无标签扩展策略的基础上,将考虑了数据不平衡的错分代价融入到 Laplacian 支持向量机[14]的经验损失和 Laplacian 正则化项中.同时,为了考虑噪声样本对决策平面的影响,本文定义了一种样本依赖的代价,对噪声样本赋予较低的权重.本文在 UCI 数据集和 NASA 软件数据集上对本文算法的性能进行了验证.实验结果表明,在 UCI 数据集和 NASA 软件数据集上,本文算法都取得了较低的错分代价,充分体现了本文算法的有效性.

2　相关工作

2.1　代价敏感的半监督支持向量机

设 $S=\{(x_i,y_i)\mid i=1,\cdots,n\}\subset R^d\times\{+1,-1\}$ 为给定的训练集,x_i 表示第 i 个训练样本,y_i 为该样本的类别.假设 S 中有 l 个有标签样本对 $\{(x_1,y_1),\cdots,(x_l,y_l)\}$ 和 u 个无标签样本对 $\{(x_{l+1},y_{l+1}),\cdots,(x_{l+u},y_{l+u})\}$,对应的索引分别为 I_l,I_u.正类错分成负类的代价为 C^+,负类错分成正类的代价为 C^-.

假设无标签样本的标签 $\hat{y}=[\hat{y}_i;i\in I_u]$,将错分代价融入支持向量机(SVM)的经验损失中,可以得到代价敏感的半监督支持向量机:

$$\min_{\hat{y}\in B}\min_{f}\frac{1}{2}\|f\|^2+C_1\sum_{i\in I_l}l(y_i,f(x_i))+C_2\sum_{i\in I_u}l(\hat{y}_i,f(x_i)) \tag{1}$$

其中,$B=\{\hat{y}\mid\hat{y}_i\in\{\pm1\},\hat{y}^Tt=r\}$ 表示无标签样本的标签的假设集合,t 是全 1 列向量,r 用于控制假设的无标签样本的标签中的正负类样本比;$l(y_i,f(x_i))=C^{y_i}\varepsilon_i$ 表示样本 x_i 的代价敏感的损失函数,C^{y_i}表示将 y_i 类样本错分的代价,$\varepsilon_i=1-y_if(x_i)$表示样本 x_i 的错分量,f 是预测函数.C_1,C_2 均为惩罚参数.

由于式(1)的求解时间复杂度高,文献[6]的作者先估计无标签样本的正负类均值,然后将样本均值引入式(1),得到一个新的代价敏感的半监督支持向量机模型(cs4vm).cs4vm 降低了时间复杂度,但没有考虑:(a)样本间的结构信息;(b)数据的不平衡性;(c)噪声问题.

2.2　半监督 Laplacian 支持向量机

基于图的学习方法[8,14,15,17]是机器学习和模式识别领域的一个研究热点.图 W 可以表示为(V,E),训练集的每一个样本都为图的一个顶点 $V=\{v_1,v_2,\cdots,v_n\}$,E 表示边的集合,具体的定义如下:

$$W_{ij}=\begin{cases}\exp(-\|x_i-x_j\|^2/\sigma^2), & \text{if } x_i\in ne(x_j)\\ & \text{or } x_j\in ne(x_i)\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

其中,$ne(x_i)$表示训练集中离样本 x_i 最近的 k 个样本集合.

将近邻图 W 引入 SVM 模型中,我们得到半监督 Laplacian 支持向量机模型[3]:

$$\begin{aligned}&\min_f:\frac{1}{l}\sum_{i=1}^{l}\varepsilon_i+r_A\|f\|_K^2+r_I\|f\|_I\\&\text{s.t.}\quad y_if(x_i)\geqslant1-\varepsilon_i,\varepsilon_i\geqslant0\end{aligned} \tag{3}$$

其中,l 是训练集中有标签样本的个数,r_A,r_I 是惩罚参数,$\|f\|_K^2$ 是核正则化项,用于控制模型 f 的复杂度.根据聚类假设:如果两个样本属于近邻关系,那么他们的标签应该是一样的,可得 $\|f\|_I=\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(f_i-f_j)^2$.

3　代价敏感的半监督 Laplacian 支持向量机

为了挖掘无标签样本的标签信息,本文首先采用文献[6]中的方法估计无标签样本的标签信息,然后在 Laplacian 支持向量机模型中考虑数据的不平衡性和噪声问题.我们可得代价敏感的半监督 Laplacian 支持向量机模型:

$$\begin{aligned}\min_f:\ &C_1\frac{1}{l}\Big(\sum_{i\in S^+}C_i^+\varepsilon_i+\sum_{i\in S^-}C_i^-\varepsilon_i\Big)+\\&C_2\frac{1}{u}\Big(\sum_{i\in S^+}C_i^+\varepsilon_i+\sum_{i\in S^-}C_i^-\varepsilon_i\Big)+r_A\|f\|_K^2+r_If^TLf\end{aligned} \tag{4}$$

其中,C_1,C_2 表示惩罚参数,S^+,S^- 表示训练集中正负类样本的索引,l,u 分别表示训练集中有标签样本和无标签样本的个数.$C_i^+=w_i\dfrac{C^+}{N^+}$表示第 i 个正类样本的错分代价,$w_i=N^+d_i\Big/\sum_{j=1}^{N^+}d_j$,$d_i=\sum_{j=1}^{N^+}W_{ij}$;$N^+,N^-$ 分别表示训练集中正负类样本个数,W 是近邻图,定义如式(2).$\boldsymbol{L}=\begin{bmatrix}\dfrac{C^+}{N^+}\cdot\boldsymbol{L}^+ & \\ & \dfrac{C^-}{N^-}\cdot\boldsymbol{L}^-\end{bmatrix}$,$\boldsymbol{L}^+,\boldsymbol{L}^-$ 分别表示正负类样本的 Laplacian 矩阵[14].

噪声样本会影响 SVM 的分类平面,尤其当噪声样本在分类边界上,所以我们希望给噪声样本赋予一个较低的权重.为此,本文提出了 $w_i=N^+d_i\Big/\sum_{j=1}^{N^+}d_j$ 的加权策略.如果 w_i 值越大,那么样本 x_i 周围的样本越稠密,x_i 为噪声的可能性就比较低;如果 w_i 值越小,那么样本 x_i 周围的样本越稀疏,x_i 为噪声的可能性就比较

高.又由于,w_i 依赖于局部近邻图 W,所以 w_i 不依赖于数据分布,不仅能够适应于高斯分布的数据,也能够应用于流形分布的数据.

根据文献[4,5]的作者提出的将数据不平衡性和代价统一考虑的学习框架,可得新的正负类错分代价 C^+/N^+,C^-/N^-.综合考虑数据不平衡性和噪声样本,本文提出了一种样本依赖的错分代价 $C_i^+ = w_i C^+/N^+$.

通过标签扩展,我们可得无标签样本的标签信息.为此可以定义监督的 $\boldsymbol{W} = \begin{bmatrix} W^+ & 0 \\ 0 & W^- \end{bmatrix}$,$W^+$,$W^-$ 分别表示正负类样本的近邻图.正类样本的代价敏感的局部结构保持定义为:$\sum_{i=1}^{N^+}\sum_{j=1}^{N^+}\frac{C^+}{N^+}(f_i - f_j)^2 W_{ij}^+$.这里,我们把$(f_i - f_j)^2$ 理解为第 i 个正类样本的错分量.通过化简,我们不难得到 $f^T\frac{C^+}{N^+}L^+ f$.

如果不考虑数据的不平衡性以及噪声问题,即令 $N^+ = N^-$,$w_i = 1$,那么式(4)就退化为仅将错分代价融入到经验损失中的代价敏感的半监 Laplacian 支持向量机(cs-lapsvm);如果令 $N^+ = N^-$,$w_i = 1$,$r_I = 0$,那么式(4)退化为代价敏感的支持向量机[7].

4 实验

4.1 数据集描述

本文选择 8 个较平衡的 UCI[11]数据集作为平衡数据集,8 个 NASA[12]数据集作为不平衡的数据集,详细描述见表 1.其中,Heart1,Heart2,Ion 分别表示 Heart_disease,Heart_statlog,Ionosphere 数据集.

表 1 数据集的详细描述

UCI 数据集				NASA 数据集			
数据集	正类样本数	负类样本数	属性个数	数据集	正类样本数	负类样本数	属性个数
Heart1	150	120	13	CM1	48	456	36
House	108	124	16	KC3	43	386	39
Ion	225	126	33	MC1	46	1960	38
Sonar	97	111	60	MW1	31	348	37
Wdbc	148	46	32	PC1	69	886	37
Heart2	150	120	13	PC3	153	1283	37
Musk	207	269	166	PC4	177	1166	37
Kr-vs-kp	1669	1527	36	PC5	484	1333	38

4.2 实验参数设置

在半监督学习中,训练集中的监督信息很少,因此很难用交叉验证的方法来进行参数选择;同时,考虑到本文模型是 Laplacian SVM 模型的推广,为了比较的公平性,本文采用文献[6,14]中 Laplacian SVM 的参数设置:$r_A = 1$,$r_I = 0.1$,$C_1 = 1$.本文需要扩展无标签样本的标签信息,因此无标签样本的预测准确率,对本文模型的性能起着重要的影响,这里,我们通过参数 C_2 加以控制.通过对 C_2 进行参数选择,可得 $C_2 = 0.3$.如果不进行标签扩展 $C_2 = 0$,那么本文参数 r_A,r_I,C_1 的设置和文献[6,14]中 Laplacian SVM 的参数设置一致;如果令 $r_I = 0$,Laplacian SVM 模型退化为 SVM 模型,那么本文参数 r_A,C_1 的设置和文献[6]中 cs4vm 的一致.

4.3 实验结果分析

在实验中,我们以测试集上的整体错分代价作为评价标准,分别与以下 4 个算法进行比较:(1)仅采用 l 个有标签样本进行训练的代价敏感的支持向量机(cs-svm)[7];(2)将错分代价融入到 l 个有标签样本的经验损失中的代价敏感的半监督 Laplacian 支持向量机(cs-lapsvm);(3)代价敏感的半监督支持向量机(cs4vm)[6];(4)采用整个训练集进行训练的代价敏感的支持向量机(gt-svm)[7].

4.3.1 UCI 数据集上的算法性能比较

为了检验在平衡数据集上,错分代价的变化对算法性能的影响.本文令 $C^- = 1$,C^+ 分别为 5,10,50.UCI 数据集每类的一半作为训练集,每类中任取 5 个样本作为监督样本.实验结果见图 1.

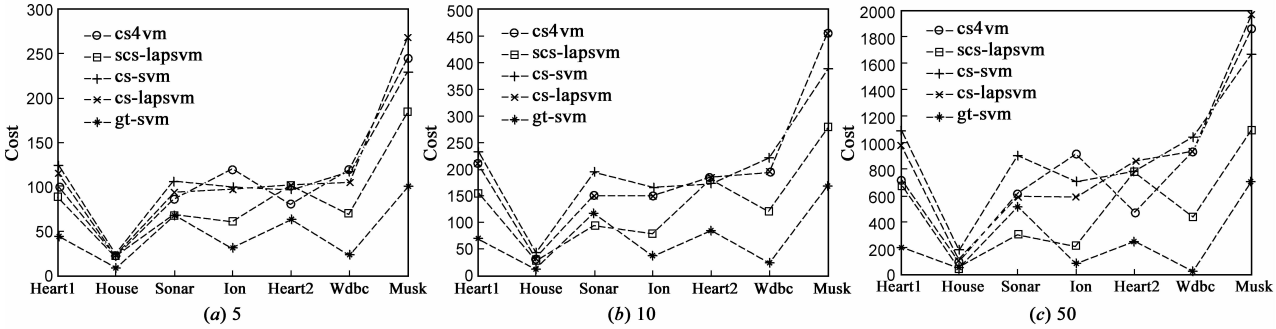
为了检验在平衡数据集上,监督信息对算法性能的影响.本文令 $C^+ = 10$,$C^- = 1$,训练集中每类的监督样本分别为 10,15,20.实验结果见图 2.

从图 1,2 中,我们观察发现:在 UCI 数据集上,本文模型 scs-lapsvm 和 cs-svm,cs-lapsvm,cs4vm 相比,在不同的监督样本和正负类错分代价比下都取得了较好的结果.正负类错分代价比越大,scs-lapsvm 的优势越明显.当每类监督样本为 20 的时候,scs-lapsvm 在 Sonar 和 Musk 数据集上的错分代价低于 gt-svm.可见,随着监督信息的增加,scs-lapsvm 的性能有可能接近或者超越 gt-svm.

4.3.2 NASA 软件数据集上的算法性能比较

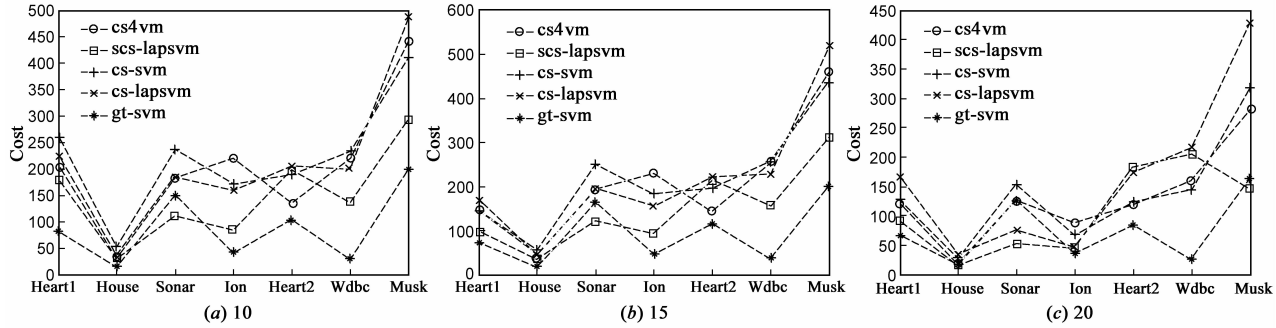
在不平衡的 NASA 软件数据集上,为了检验监督信息对算法性能的影响,本文随机选取 300 个样本作为训练集,监督样本分别为 50,100,150.$C^+ = 70$,$C^- = 1$.采用 95% 置性度水平下的 t 测试对实验结果进行显著性检验.实验结果见表 2.从表 2 中,我们观察发现:scs-lapsvm 的错分代价明显低于 cs-svm 和 cs4vm,这是因为 scs-lapsvm 考虑了数据的不平衡性.cs-lapsvm 和 scs-lapsvm 的性能是可比较的,这说明在 NASA 软件数据集上采用 Laplacian 的聚类假设是合理的.随着监督样本的增加,scs-lapsvm 的性能逐渐优于 gt-svm,当监督的样本为 150 的时候,scs-lapsvm 在 8 个数据集上都优于 gt-svm.

为了研究数据不平衡性和代价之间的关系,本文在 NASA 数据集上,固定监督的样本数为 50,$C^- = 1$,C^+分别为5,10,30,50,70,90,120.实验结果见图3.从



(*a*) 5　(*b*) 10　(*c*) 50

图1　算法在不同正负类错分代价比下的性能比较.正负类样本错分代价比分别为:(*a*) 5;(*b*) 10;(*c*) 50.



(*a*) 10　(*b*) 15　(*c*) 20

图2　算法在不同监督样本数下的性能比较. 每类的监督样本数分别为:(*a*) 10;(*b*) 15;(*c*) 20.

表 2　在 NASA 软件数据上不同监督样本数下的整体错分代价比较(平均值 ± 标准差)

监督样本数	数据集	cs-svm	cs-lapsvm	cs4vm	scs-lapsvm	gt_svm
50	CM1	976.6 ± 221.4	306.7 ± 287.9	1014.36 ± 196.54	275.86 ± 235.7	396.6 ± 152.10
	KC3	616.86 ± 154.7	163.2 ± 61.08	631.9 ± 152.03	223.5 ± 212.6	287.16 ± 125.12
	MC1	2595.5 ± 183.8	1850.86 ± 280.61	2612.7 ± 184.41	2105.9 ± 431.3	2089.53 ± 310.8
	MW1	290.53 ± 102.98	113.4 ± 82.3	281.23 ± 92.07	160.73 ± 130.07	169.1 ± 80.99
	PC1	2483.36 ± 369.37	971.6 ± 685.55	2475.96 ± 336.9	1377.96 ± 1066.18	1019.8 ± 372.7
	PC3	4337.5 ± 1262.13	1206.53 ± 567.81	4315.06 ± 984.01	1567.03 ± 1474.1	1801.96 ± 453.88
	PC4	3334.6 ± 1594.6	1817.4 ± 2473.2	3440.36 ± 1215.86	1164.5 ± 1365.39	1062.46 ± 300.1
	PC5	6202.03 ± 2196.63	14006.97 ± 11412.36	5799.533 ± 2813.9	1140.6 ± 54.62	1280.8 ± 354.56
scs-lapsvm: W/T/L		8/0/0	1/5/2	8/0/0		2/5/1
100	CM1	674.23 ± 198.48	243.83 ± 154.67	681.5 ± 177.02	256.53 ± 201.07	344.2 ± 142.5
	KC3	609.63 ± 144.51	203.7 ± 157.26	607.26 ± 139.3	171.3 ± 149.54	313.3 ± 114.98
	MC1	2448.26 ± 208.7	1762.73 ± 246.38	2485.86 ± 202.82	1970.26 ± 429.14	2124.53 ± 278.21
	MW1	2581.33 ± 115.87	87.43 ± 3.16	276.3 ± 110.23	111.33 ± 87.63	170.06 ± 71.73
	PC1	1807.26 ± 419.72	1056.66 ± 834.53	1833.8 ± 442.45	865.13 ± 513.38	891.53 ± 270.13
	PC3	3501.93 ± 1011.9	1171.96 ± 506.22	3484.3 ± 1071.65	1179.7 ± 816.27	1694.8 ± 527.79
	PC4	2301.76 ± 583.50	1057.13 ± 395.71	2460.13 ± 634.15	894.9 ± 34.34	1097.73 ± 312.94
	PC5	3746.5 ± 1387.70	20694.7 ± 10353.18	2661.73 ± 1640.88	1141.6 ± 46.2	1365.43 ± 528.88
scs-lapsvm: W/T/L		8/0/0	2/5/1	8/0/0		6/2/0
150	CM1	541.43 ± 153.13	209.2 ± 39.21	519.3 ± 152.65	188.3 ± 15.61	352.7 ± 152.41
	KC3	485.16 ± 117.04	185.1 ± 139.50	468.96 ± 124.12	126.43 ± 63.99	329.1 ± 127.46
	MC1	2233.36 ± 271.08	1716.76 ± 96.55	2286.46 ± 233.31	1818.6 ± 359.58	2033.3 ± 257.40
	MW1	234.93 ± 101.10	82.9 ± 23.87	239.4 ± 103.73	85 ± 54.36	178.8 ± 76.31
	PC1	1402 ± 395.43	885.73 ± 606.24	1434.9 ± 438.90	645.4 ± 64.92	837.36 ± 272.75
	PC3	2490.06 ± 651.93	1057.8 ± 198.8	2385.5 ± 799.44	1009.26 ± 0.62	1722 ± 629.37
	PC4	1795.26 ± 518.30	1233.8 ± 1557.71	1819.16 ± 592.95	905.56 ± 1.05	1034.26 ± 319.43
	PC5	2404.46 ± 787.09	16534.80 ± 11204.12	1827.5 ± 783.49	1113.6 ± 0.48	1366.73 ± 435.64
scs-lapsvm: W/T/L		8/0/0	4/4/0	8/0/0		8/0/0

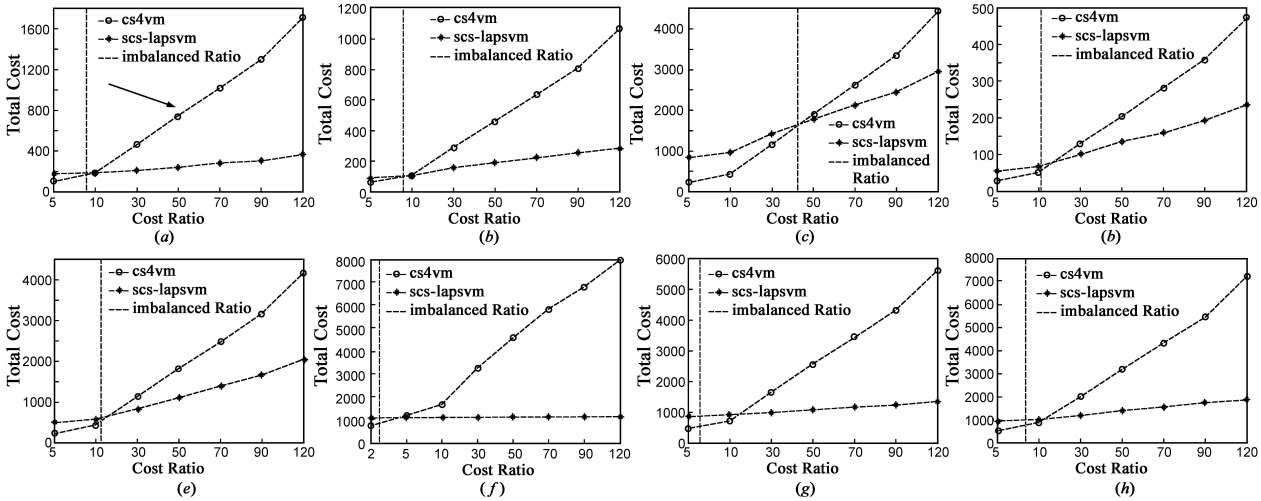


图3 scs-lapsvm和cs4vm在NASA软件数据集上不同正负类样本错分代价比下的性能比较.
图(*a-h*)分别表示NASA数据集CM1,KC3,MC1, MW1, PC1, PC3, PC4, PC5.

图3中,我们观察发现:当样本的错分代价比大于数据的正负类不平衡率时,scs-lapsvm的错分代价低于cs4vm,因此需要考虑数据不平衡性;当样本的错分代价比小于数据的正负类不平衡率时,cs4vm的错分代价要明显低于scs-lapsvm,因此只需要考虑错分代价.本文的这一结论也正好验证了文献[5]中的结论.

4.3.3 参数 C_2 的选择

本文通过变化 C_2 的值,检验 C_2 对 scs-lapsvm 性能的影响.在实验中,监督的样本数为 50, C_2 的值分别为:0,0.005,0.01,0.03,0.05,0.06,0.07,0.08,0.1,0.5,1.从图4中,我们观察发现 C_2 取[0.03,0.06]比较合适.

4.3.4 时间复杂度分析

本文模型 scs-lapsvm 的时间复杂度分为两个部分:(1)无标签扩展的时间复杂度,这和 cs4vm 无标签扩展的时间复杂度一致;(2)求解 scs-lapsvm 模型的时间复杂度,scs-lapsvm 模型是 Laplacian SVM 模型的扩展,两者的时间复杂度均为 $O((l+u)^3)$[14].算法的时间复杂度比较见表3.从表3,我们可知:本文模型 scs-lapsvm 在中小规模的数据集上的时间复杂度略高于 cs4vm;在大规模数据集上的求解效率有所下降.

表3 算法的时间复杂度比较(单位:s)

数据集(样本数)	Sonar (208)	Ino (351)	Musk (476)	Pc1 (957)	Pc4 (1343)	Pc5 (1817)	Kr-vs-kp (3196)
cs-lapsvm	0.1016	0.1151	0.1276	0.1802	0.1818	0.1995	3.51
cs4vm	0.2318	0.2391	0.2984	1.3042	1.333	1.6974	1.5516
scs-lapsvm	0.2433	0.2615	0.3176	1.3771	1.3948	1.7776	3.9704

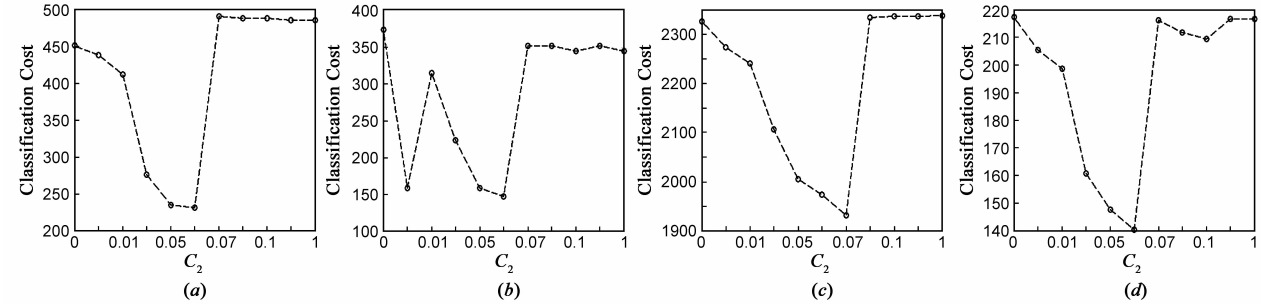


图4 C_2取不同值时scs-lapsvm的错分代价.图(*a-d*)分别表示NASA数据集CM1,KC3,MC1,MW1.

5 总结和展望

本文针对数据集的不平衡性,存在大量的无标签样本,只有少量的有标签样本,并且存在噪声样本,提出了一种代价敏感的半监督 Laplacian 支持向量机(scs-lapsvm).实验在7个 UCI 数据集和8个 NASA 软件数据集上,分别对 scs-lapsvm 进行了验证.实验结果表明了本文算法的有效性.

本文算法取得了较好的实验结果,但仍有一些问题值得我们继续去研究:(1)提升算法在大规模数据集下的求解效率;(2)自适应的学习不平衡数据集上的错分代价.

参考文献

[1] R Batuwita, V Palade. FSVM-CIL: Fuzzy support vector machines for class imbalance learning [J]. IEEE Transactions on Fuzzy Systems, 2010, 18(3): 558 – 571.

[2] U Brefeld, P Geibel, et al. Support vector machines with example dependent costs [A]. Proceedings of the European Conference on Machine Learning [C]. Gavtat-Dubrovnik, Croatia, 2003. 23 – 34.

[3] N V Chawla, N Japkowicz, et al. Editorial to the special issue on learning from imbalanced data set[J]. ACM SIGKDD Explorations, 2004, 6(1): 1 – 6.

[4] C Elkan. The foundations of cost-sensitive learning [A]. Proceedings of the 17th International Joint Conference on Artificial Intelligence[C]. San Francisco, CA, USA, 2001. 973 – 978.

[5] X Y Liu, Z H Zhou. The influence of class imbalance on cost-sensitive learning: An empirical study[A]. Proceedings of the 6th IEEE International Conference on Data Mining[C]. Hong Kong, China, 2006. 970 – 974.

[6] Y F Li, J Kwok, et al. Cost-sensitive semi-supervised support vector machine[A]. Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10) [C]. Atlanta, GA, 2010. 500 – 505.

[7] K Morik, P Brochhausen, et al. Combining statistical learning with a knowledge-based approach: A case study in intensive care monitoring [A]. Proceedings of 16th International Conference on Machine Learning [C]. San Francisco, CA, USA, 1999. 268 – 277.

[8] L Qiao, S Chen, et al. Sparsity preserving discriminant analysis for single training image face recognition [J]. Pattern Recognition Letters, 2010, 31(5): 422 – 429.

[9] Y Zhang, Z H Zhou. Cost-sensitive face recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(10): 1758 – 1769.

[10] Z H Zhou, X Y Liu. On multi-class cost-sensitive learning [J]. Computational Intelligence, 2010, 26(3): 232 – 257.

[11] C Blake et al. UCI repository of machine learning databases [DB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998-04-02.

[12] M Chapman, P Callis, et al. Metrics data program [DB/OL]. http://mdp.ivv.nasa.gov, 2004.

[13] K Bennett, A Demiriz. Semi-supervised support vector machines[A]. Advances in Neural Information Processing Systems 11[C]. MIT Press, 1999. 368 – 374.

[14] M Belkin, P Niyogi, et al. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples [J]. Journal of Machine Learning Research, 2006, 7(48): 2399 – 2434.

[15] D Cai, X He, et al. Semi-supervised discriminant analysis [A]. IEEE International Conference on Computer Vision [C]. Rio de Janeiro, Brazil, 2007. 1 – 7.

[16] O Chapelle, B Scholkopf, et al. Semi-Supervised Learning [M]. Cabridge, MA: MIT Press, 2006.

[17] S Roweis, L Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290: 2323 – 2326.

作者简介

万建武　男,1986 年出生于江苏常州.现为南京师范大学应用数学专业硕博连读生.从事机器学习,模式识别方面的相关研究工作.

E-mail: xiaowunjnu@163.com

杨　明　男,1964 年 11 月生,博士,教授,博士生导师,安徽宁国人.主要研究领域为数据挖掘,机器学习,模式识别等.