安全意识的半监督分类

汪云云，陈松灿

***摘*要 - 虽然半监督分类学习在过去几十年中引起了极大的关注，但在某些情况下，半监督分类方法的表现可能比有监督的对应方法更差，从而降低了他们对实际应用的信心。 当然，期望开发一种安全的半监督分类方法，其永远不会比受监督的对应方更差。 然而，就我们所知，很少有研究致力于安全的半监督分类。 为了解决这个问题，在本文中，我们通过半无监督数据和监督分类之间的自适应权衡，发明了安全半监督分类的安全控制机制。 在实施中，基于我们最近基于类成员资格（SSCCM）的半监督分类方法，我们开发了安全感知SSCCM（SA-SSCCM）。 SA-SSCCM一方面利用未标记的数据来帮助学习（如SSCCM所做的那样），假设未标记的数据可以帮助学习，另一方面，限制其预测接近其受监督的对应最小二乘的预测支持向量机（LS-SVM）假设未标记的数据可能会伤害学习。 因此，就未标记数据而言，SA-SSCCM的预测分别在半监督SSCCM和监督LS-SVM之间进行权衡。 与SSCCM一样，SA-SSCCM中的优化问题可以通过交替迭代策略有效地解决，理论上可以保证迭代收敛。 对几个真实数据集的实验表明，与LS-SVM，SSCCM和现成的安全半监督分类方法相比，SA-SSCCM具有良好的性能。**

***索*引术语 - 交替迭代策略，最小二乘支持向量机（LS-SVM），半监督分类，基于类成员的半监督分类（SSCCM）。**

1. 介绍

**I**

在许多实际应用中，可以容易且廉价地收集未标记的数据，而标记数据的获取通常非常昂贵且耗时，尤其涉及手动操作。 例如，在网页推荐中，可以获得大量的网页，但很少有用户愿意花费大量时间来标记他们感兴趣的网页。在垃圾邮件检测中，可以自动收集大量电子邮件，但很少

手稿于2012年8月8日收到; 2013年5月5日接受。日期

出版物2013年6月17日; 当前版本的日期为2013年10月15日。这项工作部分得到了国家科学基金会的资助61035003，格兰特61170151，格兰特61272422和格兰特61073114，以及江苏青兰项目。

Y. Wang在南京邮电大学计算机科学与工程系，南京210046（电子邮件：wangyunyun@njupt.edu.cn）。

S. Chen在南京航空航天大学计算机科学与工程系，南京210016（电子邮件：s.chen@nuaa.edu.cn）。

本文中一个或多个图的彩色版本可在线获取 [http://ieeexplore.ieee.org。](http://ieeexplore.ieee.org/)

数字对象标识符10.1109 / TNNLS.2013.2263512

他们被标记为垃圾邮件或用户没有。 因此，半监督学习在过去的几十年中引起了人们的广泛关注，这种学习利用大量未标记的数据和有限的标记数据进行学习。 在本文中，我们关注半监督分类，并且在过去几十年中开发了许多半监督分类方法[1] - [4]。

半监督分类方法试图利用未标记数据所公开的内在数据分布信息，并且通常认为该信息有助于学习。 为了利用未标记的数据，应该采用一些假设进行学习。 半监督分类中的两个常见假设是聚类假设和多重假设[3] - [5]。 群集假设假设类似的实例可能共享相同的类别标签，从而引导通过低密度区域的分类边界。 多重假设假设数据分布在由拉普拉斯图表示的某些低维流形上，并且类似实例应根据图形共享相似的分类输出。 几乎所有现成的半监督分类方法都明确地或隐含地采用这些假设中的一个或两个[1]，[4]。 例如，大边缘半监督分类方法，如转导支持向量机（TSVM）[6]，半监督SVM（S3VM）[7]及其变体[8]，[9]，采用聚类假设。 基于图的半监督分类方法，如标签传播[10]，[11]，图形切割[12]和拉普拉斯SVM（LapSVM）[13]，采用多种假设。 这些半监督分类方法应用于各种应用，例如文本分类[6]，垃圾邮件检测[14]和生物信息学[15]等。 然而，在某些情况下，由于使用了未标记的数据，半监督分类方法可能会产生比监督对应方更差的性能[16] - [18]。 由于模型假设或数据分布假设的失败，未标记的数据可能会损害性能或对学习有害[4]，[16]，[19]。 考虑到这一点，Wang等人。 [19]提出了一个修改的集群假设，即类似的实例应该共享相似的类成员资格而不是一个清晰的类标签，因此开发了一种基于类成员资格（SSCCM）的新的半监督分类方法。 修正的聚类假设可以比聚类假设更好地捕获实际数据分布，因此，SSCCM在很大程度上实现了比基于聚类假设的半监督分类方法更好的性能。 然而，与其他半监督分类方法一样，SSCCM在某些情况下仍然比其受监督的对应最小二乘（LS）-SVM表现更差。 它降低了申请的信心

2162-237X © 2013 IEEE

实际应用的半监督方法。 因此，需要开发一种安全的半监督分类方法，该方法从未比其受监督的对应物表现更差，或者由于使用未标记的数据而导致性能退化[17]。 据我们所知，到目前为止，最近只有两项关于安全半监督分类的研究[17]，[18]。 李等人。 [18]通过层次聚类选择未标记的实例来开发S3VM-us，以便仅利用非常可能有用的未标记实例。 更具体地说，在S3VM-us中，TSVM仅预测通过层次聚类具有高置信度的未标记实例，其余由SVM预测。 最后，其性能退化的可能性远小于TSVM。 同时，李等人。

[17]还开发了安全的S3VM（S4VM）方法。 与寻求最佳低密度分离器的S3VM不同，S4VM同时利用候选低密度分离器，以降低识别带有未标记数据的差分离器的风险。 具体而言，S4VM找到多种大边距低密度分离器，然后在最坏情况下优化未标记数据的标签分配，假设可以通过获得的低密度分离器之一实现地面实况标签分配。 最后，S4VM的性能与TSVM相比具有很强的竞争力，并且从未显着低于（监督的）SVM。 S3VM\_us和S4VM实际上都是基于归纳S3VM的改进，同时属于转换式方法，它学习给定的标记和未标记数据，并且仅针对给定的未标记数据寻找标签。

当可用的未标记数据对于学习是不可靠的时，预期安全的半监督分类方法不会比监督对应方差，并且当可用的未标记数据对于学习可靠时，不比半监督方法差。 简单地说，当给出不可靠的未标记数据时，我们期望最终预测接近受监督对应物的预测，并且当给出可靠的未标记数据时，接近半监督方法的预测。 为此，本文通过自适应发明了一种安全半监督分类的安全控制机制

控制半监督和超监督之间的权衡

从来没有明显低于LS-SVM，并且很少明显低于SSCCM。

S3VM\_us和S4VMs的开发都不比SVM差，而SA-SSCCM的开发并不比监督和半监督对应物差，这反映在实证结果中。 此外，与转换S3VM\_us和S4VM不同，SA-SSCCM以归纳方式学习，其在给定的标记和未标记数据上进行学习，并且旨在寻找用于预测看不见的实例的判别函数。 值得指出的是，尽管我们在本文中涉及SSCCM，但所提出的机制可以很容易地应用于其他半监督方法，如TSVM，S3VM和LapSVM。

本文的其余部分安排如下。 第二节简要介绍了SSCCM，第三节描述了拟议的SA-SSCCM，第四节描述了实证结果，最后，结论总结在第五节。

1. 基于类成员的半监督分类方法

集群假设假设类似的实例应该共享相同的类标签，因此隐含地假设每个实例应该具有清晰的标签分配。 但是，在实际应用程序中，例如边界实例，很难将实例分配给单个类。 在这些情况下，群集假设不能充分反映实际数据分布，并且在半监督分类中采用时会导致较差的预测。 因此，王等人。 [19]建议修改集群假设，假设类似的实例应该共享相似的类成员资格而不是一个清晰的类标签。 对于每个实例，类成员资格表示为向量，每个元素的值表示相关实例属于该类的可能性。 通过采用修改的聚类假设以及局部学习原则（约束每个实例及其局部加权均值（LWM）共享相同的标签隶属度向量），Wang等。 [19]开发了一种名为SSCCM的新型半监督分类方法。

现在，给定标记数据X*l*= {x*i*}*吧*与对应 -

标签Y = {y}*吧*

*i*=1

*X* = {*x* }*n*

根据可用的未标记数据进行分类。

1. *i=1*和未标记的数据 *u*
2. *j* =*nl*+1

在实施中，基于我们最近的SSCCM，我们开发了安全感知SSCCM（SA-SSCCM）。 SA-SSCCM一方面利用未标记的数据来帮助学习（如SSCCM所做的那样），假设未标记的数据

= −

∈ ∈

其中每个x*i*R*d*，y*i*R*C*用于C类分类，和n*u*nn*l*。 利用决策函数f（x）和标签隶属函数v（x），SSCCM可以表述如下：

另一方面，可以帮助学习将其预测限制在 *吧 吧*

接近其监督对应的LS-SVM 分 || f（xi） - yi ||2+λs|| f（xi） - yi ||2

假设未标记的数据可能会损害学习。 最后，SA-SSCCM的预测是在那些之间的权衡

*f，vk（xj）*

*i*=1

*C n*

*i*=1

SSCCM和LS-SVM分别就未标记而言

数据。 此外，我们提供了一种调整此类权衡的策略

就给定的未标记数据而言。 与SSCCM一样，SA-SSCCM中的优化问题可以通过交替迭代策略有效地解决，并且理论上可以保证迭代的收敛。 几个真实数据集的实证结果表明，SA-SSCCM的整体性能优于SSCCM和LS-SVM，并且SA-SSCCM的性能也是如此。

+ *vk（xj）*2*|| f（xj） - rk||*2

*k*=1 *j* =*nl* +1

+λ*s*v*k*（x*j*）2|| f（x*j*） - r*k*||2+λ|| f ||2

*C*

*n*

*k*=1 *j* =*nl* +1 *H*

*C*

S*T vk（xj）= 1*

*k*=1

0*≤vk（xj）≤1，k = 1 ... C， j* = n*l*+ 1 ... n（1）

其中||·||*H*是再生核Hilbert空间中的一个范数，

未标记的数据有助于学习和方法

*C*

{*r}k*

∈

和*k* = 1时

是C类的编码，两者都是*i RC*

*C*

∈

当未标记的数据是监督的LS-SVM时

*Rk R* 由一个C规则编码，即第k个

如果x*i*属于第k类，则y*i*的元素为1，其余为零，r*k*的第k个元素为1，其余为零，f（x*i*）R*C* ）和v（x*i*）R*C*对于每个x*i*，并且v*k*（x*i*）表示x*i*属于第k类的可能性，x*i*是LWM

ˆ

∈ ∈

由x定义的*i*

对学习有害。 使用LS-SVM的预测

学习，我们将从LS-SVM得到的决策函数表示为g（x），然后用决策函数f（x）和标签隶属函数v（x），我们建立SA-SSCCM如下：

*n*

*H*

*x* ,*xj∈Ne（xi）Sijxj*



分

*f，vk（xj）*

|| *f* ||2

+λ1|| f（x*i*） - y*i*||2

ˆ*i* = ,

*xj∈Ne（xi）*

*骶髂关节* (2)

*i*=1

*C n*

I

*k*=1 *j* =*nl* +1

其中Ne（x*i*）是测量的x*i*的k最近邻域

通过欧几里德距离，Sij是一个反向的数量

+*λ*2

*vk（xj）*2*|| f（xj） - rk||*2

通过欧几里德距离，Sij是一个反向的数量

特定于x*i*和x*j*之间的距离，∀k= 1 ... C，i =

···

1 *n*.

在SSCCM中，每个实例都可以属于具有相应类成员身份的多个类，此外，

1

+ *λ* − 1

\

*C*

*n*

*j* = *nl* +1

|*| f（xj） - g（xj）||*2

每个实例及其LWM共享相同的标签成员资格向量。 优化问题可以通过交替迭代策略有效地解决，其中每个步骤都生成

S*T*

*vk（xj）= 1*

*k*=1

0*≤vk（xj）≤1，k = 1 ... C，j = nl+ 1 ... n。 （4）*

封闭形式的解决方案。 迭代求解过程的收敛在理论上可以得到保证。 最后，与几种最先进的半监督分类方法（如TSVM，LapSVM和mean3svm）相比，SSCCM实现了竞争性能。

与其他半监督分类方法一样，SSCCM可以产生比其受监督的对应LS-SVM更差的性能。 因此，在本文中，我们试图进一步开发一种安全感知的半监督分类方法。 然而，在SSCCM中的每个LWM的计算中，存在潜在风险[20]，来自相对类的实例也可以在k个最近邻域中被选择，因此所获得的LWM可以落入相反的类中。 为了避免在SA-SSCCM建模开始时存在这样的风险，我们从（1）中删除涉及LWM的术语并简化SSCCM的优化问题（为方便起见，我们在下文中滥用名称SSCCM）如下：

*吧*

*H*

（4）目标函数中的前三项通过像SSCCM那样利用标记和未标记数据同时寻找决策函数f（x）和标签隶属函数v（x），而最后一项控制预测之间的差异（对于未标记的实例）分别由SA-SSCCM和LS-SVM。 因此，SA-SSCCM的预测分别成为SSCCM和LS-SVM之间的折衷，由λ调整。 λ根据可用的未标记数据从[0,1]获取其值，当λ接近零时（当λ接近零时，令1 /λ为无穷大），SA-SSCCM的预测退化为LS-SVM的预测，并且当λ接近1，SA-SSCCM的预测退化为SSCCM的预测。 因此，λ在控制SSCCM和LS-SVM之间权衡的SA-SSCCM中起着重要作用，其值将在后面的第III-D节中讨论。

分

*f，vk（xj）*

|| *f* ||2

+ *λ*1 || f（x*i*） - y*i*||2

*i*=1

*B.优化*

与SSCCM一样，SA-SSCCM中的优化问题是

*C*

+*λ*2

*n*

*vk（xj）*2*|| f（xj） - rk||*2

（f，v）中的biconvex，我们可以求助于具有收敛保证的交替迭代求解策略[21]。

S*T*

*k*=1 *j* =*nl* +1

*C*

*vk（xj）= 1*

*k*=1

优化迭代中的每个步骤都为f（x）和v（x）生成闭合形式的解。

对于固定的v（x），SA-SSCCM的优化问题可以重铸如下：

0≤v*k*（x*j*）≤1，k = 1 ... C， *j* = n*l*+ 1 ... n。 （3）

1. 安全意识 SSCCM

分

*f，vk（xj）*

|| *f* ||2

+ *λ*1

*吧*

*i*=1

|| f（x*i*） - y*i*||2

在这里，我们描述了提出的SA-SSCCM方法，包括公式，优化，算法描述 -

*H*

分离部分中的参数λ和值。

*C*

*n*

+*λ*2*vk（xj）*2*|| f（xj） - rk||*2

I

\

*k*=1 *j* =*nl* +1

分离部分中的参数λ和值。

*A.配方*

通过（简化）SSCCM，我们开发了一个安全感知的半

*i*=1

1

+ *λ* − 1

*n*

*j* = *nl* +1

|*| f（xj） - g（xj）||*2*。* (5)

监督分类方法SA-SSCCM，其最终预测接近半监督SSCCM时的方法

基于代表定理[13]，其中每个α*i*

（5）的最小化符号的形式为f（x）=,n

*αiK（xi，x）*

∈ R*C*×1.

然后，（5）可以重新表述如下：

m*in M = tr（αKαT）+λtr*(*（αK - Y）（αK*

*C*

- *Y）T*\

从而

*vk（xj）=λj/ 2 || f（xj） - rk||*2*。* (12)

*α* 1 1 *l l*

*C*

此外，积分约束,Cv（x）= 1，我们

+*λ*2*tr（（αKu - Lk）V*ˆ*k（αKu - Lk）T）*

I

\

有以下内容：

*k*=1 *k j*

1 / || f（x*j*） - r*k*||2



*k*=1



*vk（xj）=，C*

(13)

2

1

+ *λ* − 1

*tr（（αKu - α*0*Ku）（αKu - α*0*Ku）T*

*)*

(6)

*k=1*1 */ || f（xj） - rk||*

其中k1 *... C , i* 1 ... n。 因此，对于任意实例x

∈ { } ∈ { }

其中α= [α1，α2，...，α*n*]∈R*C×n*是Lagrange mul-

1 l

1 / || f（x） - r*k*||2



tiplier矩阵，α0是拉格朗日乘数矩阵

*K*

*u*

*K*

*鲁*

是一个核矩阵，

*vk（x)=，C*

(14)

1 / || f（x） - r ||2

其中k∈{1 ... C}。

*k*=1

*k*

*库尔库*

= ( ± = ( ±*H H*

LS-SVM。 K = [K*l*K*u*] =

其中K*u*φ（X*l*），φ（X*l*），K*鲁*φ（X*l*），φ（X*u*）和K*UU*φ（X*u* ）），φ（X*u*）。 L*k*是C n*u*矩阵，第k行是全1矢量，而休息是全零矢量。 设V*k*为对应的标签隶属向量

= ( ± ×*H*

到第k类，Vˆk被定义为对角矩阵

对角元素是V中条目的平方值*k*。

现在，通过将M1的导数归零为α，我们得到以下结果：

*C*

与在SSCCM中一样，SA-SSCCM中每个给定实例的预测不仅可以通过决策函数f（x）来实现，还可以通过类成员函数v（x）来实现，它反映了实例对各个类的可能性。 如下面的命题1所示，与SSCCM中这两个函数可能调用的不一致预测不同，SA-SSCCM中的两个预测始终是一致的。

*命*题1：由每个给定实例的预测

决策函数和类成员函数始终保持一致。

*∂ M*1 ˆ

=*αK+λ*1*（αKl - Y）KT+λ*2 *（αKu - Lk）VkKT*



*∂α l*

I \

∀ = /=

1

*λ* − 1

暗示f*k*（x*i*）> f*j*（x*i*），j1 *... C，j k*，在哪里

**X***k*是属于第k个集群的实例集。 而



+

*k*=1

*u*

*u 证*明：对于任意实例x*i*，其类标签由

决策函数是y*i*= max*K = 1 ...用C*f*k*（x*i*），因此x*i*∈X*k*

+

*（（αKu - α*0*Ku）KT）= 0* (7)

导致以下解决方案：

I I \

*C*



由类成员函数预测的类标签是

\ *y*〜*i*= max*K = 1 ...用C*v*k*（x*i*），因此从（14），x*i*∈X*k*意味着

∀ = /=

*α=λKTY +λKV*ˆ*L +* 1 *- 1αKKT*

|*| f（xi） - rk||*2*<|| f（xi） - rj||*2*，然后f（xi）Trk> f（xi）Trj，或*

∀ = /=

1 *l* 2

I

*C*

*λ*1*KlKT+λ*2

×

*l*

*k*=1

*k*=1

*u k k*

*λ* 0 *u*

I1 \

*u*

*KuV*ˆ*k KT+ K +*

*λ*

−1

*KuKT*

*u*

\−1

∈

*.* (8)

*u*

等价地，f*k*（x*i*）> f*j*（x*i*），j 1 ... C，j k。 因此，f（x）和v（x）对x*i*X*k*的预测条件是等价的，因此两个预测是一致的。

*C.算法描述*

对于固定的f（x），SA-SSCCM的优化问题变为

*C*

*n*

SA-SSCCM的优化遵循交替迭代策略。 与SSCCM中一样，SA-SSCCM中未标记实例的标签成员资格的初始值

分

*vk（xj）*

*vk（xj）*2*|| f（xj） - rk||*2

可以通过几种策略获得，例如随机化，一些模糊聚类技术，如FCM，或简单

S*T*

*k*=1 *j* =*nl* +1

*C*

*vk（xj）= 1*

*k*=1

0*≤vk（xj）≤1，k = 1 ... C，j = nl+ 1 ... n。 （9）*

将它们设置为全零。 当M*k*M*k*1<εM*k*1时，迭代终止，其中M*k*是第k次迭代时的目标函数值，并且ε是预定阈值。 SA-SSCCM的具体算法描述如表I所示。

使用拉格朗日乘数法，我们定义以下内容：

| − |− −

*M*2*= vk（xj）*2*|| f（xj） - rk||*2*-λj*I*vk（xj） - 1*\*。*

*C*

*n*

*C*

与SSCCM一样，SA-SSCCM优化中的交替迭代过程在理论上可以保证收敛，详细证明与[19]中的相似。

*D.λ的值*

*k*=1 *j* =*nl* +1

*k*=1

(10)

SA-SSCCM中的λ控制那些之间的预测

同样，对于每个v*k*（x*j*），k 1 ... C，j 1 ... n*u*归零M2的导数，我们得到以下结果：

∀ = =

通过SSCCM和LS-SVM分别在下文中，我们尝试从[0,1]中选择它的值，这实际上是一个数据相关的问题。 当标记的实例足够时，我们可以采用典型的参数选择策略

*∂ M*2



*∂vk（xj）*

= 2 || f（x*j*） - r*k*||2 - λ*j*= 0 (11)

例如交叉验证。 当标记的实例不足时，交叉验证无效[3]，[17] - [19]，和

表I

SA-SSCCM的算法描述



**输入** *X*l：标记数据;

*Y*l：X的标签*l;*

通过采用集合策略，将设置λ的问题转化为线性组合的优化

系数{w1，w2，...，w*m*}用于关于{λ*i*}*m*的判别函数{f*i*}*m*的集合，其中有点

*i*=1

*i*=1

*徐*：未标记的数据;

*λ*，λ1，λ2：正则化参数;

*ε*：迭代停止参数;

*o*：内核参数; Maxiter：迭代的最大数量。

**输**出f（x）：决策函数;

*v*（x）：标签隶属函数。

**程序**

初始化未标记数据的标签成员资格;

将初始目标函数值设置为无穷大，即M0 = INF; 对于k = 1 ... Maxiter

通过（8）更新α，并且通过具有α的表示定理更新f（x）; 通过（14）更新v（x）;

更新目标函数值Mk; 如果| Mk - Jk-1 | <εMk-1

中断，返回f（x）和v（x）;

如果结束



因此我们采用[22]中的集合策略。 具体来说，选择λ的问题可以简化，从λ1，λ2，...，λ*m*中选择最佳值，其中λ1 0和λ*m* 1，或等效地，从f1，f2，...，f*m*中选择最佳决策函数f，其中f1和f*m*分别是从LS-SVM和SSCCM。 从半监督学习的集合策略[22]，[23]，最优f可以表示为线性组合

{ }

{ } = =

基函数{f*i*}*m*，即f =,mω*i*f*i*，,mω*i*= 1，

≥ =0*, i* = 1 *... m*

*i m*

=

*ωi*

*i* 1

。 最后，给出

{λ*I*}= 1

*i* 1

，我们可以获得

模拟到多核学习[24]的选择

对于单个内核的最佳参数，将其转换为多个内核的线性组合系数的选择。

1. 实验

在这里，我们首先提供简单的插图，以提供我们的新SA-SSCCM如何解决安全的半监督分类问题的一些直觉。 然后，我们通过与监督的LS-SVM，半监督SSCCM和现成的安全半监督分类方法S3VM\_us [18]和S4VMs [17]进行比较，评估SA-SSCCM在几个真实数据集上的性能。

1. *说明性实验*

通过利用未标记的数据，可以预期SSCCM比监督的LS-SVM表现更好。 然而，由于模型或数据分布的错误指定，未标记的数据可能会损害性能[16] - [18]，因此与其他半监督分类方法类似，SSCCM在某些情况下也会产生比LS-SVM更差的性能。 在下文中，我们将直观地说明SA-SSCCM如何在未标记数据分别损害和有益于分类的情况下工作。

我们将bupa和电离层数据集视为此处的表示。 对于每个数据集，我们随机选择一半

培训的实例，测试的休息和

对应{f}*m*

=

*i*=1

训练数据集，我们随机选择十个标记的实例，和

*ii 1*，然后将问题转换为寻求一组组合权重ω1，ω2，...，ω*m*，具体如下：

{ }

留下没有标签的休息。 这里采用线性核，正则化参数λ1，λ2,和η分别固定为100,1和1，并且ε设置为10−3。 每个边缘重量

分1



*法语 TL FR*

− -*T*

拉普拉斯图由wij= exp- || xi -xj || 2 /2σ2给出。 该

*ωk2η（*

*ω) (*

*ω)* 1*C（YlFlRω）*

未标记实例的标签成员资格的初始值

s*tωTe = 1，ωk≥0* (15)

ω[*ω*1*，ω*2*， ωm]T，e Rm×1*是所有元素等于1的向量。 F[*f*1*（X）T，f*2*（X）T... fm（X）T] Rn×（C×m）和Fl* [f1（X*l*）*T*，f2（X*l*）*T*... f*m*（X*l*）*T*] R*l×（C×m）*是整个数据集的预测矩阵和

= ∈

= ∈

= ··· ∈

分别标记数​​据集。 Y*l*∈R*C×l*是标记数据的标签矩阵。 R = [R1R2... R*m*]*T*∈R*（C×m）×m*，每个R*k*∈R*米×C*其中第k行是全1矢量且休息是全零矢量。 -CR*C ×1*是带有的载体

∈

所有元素都等于一个。 L是拉普拉斯图，表示为L. *D W*，其中W[wij]*nn*是图上的权重矩阵，D是第i个对角线条目为D的对角矩阵*ii nW*ij。 η是正则化参数。 （15）的目标函数中的第一项保证了拉普拉斯算子[13]的平滑性，第二项保证了标记数据的正确预测。 （15）的优化是二次规划（QP）问题，

= − = ×

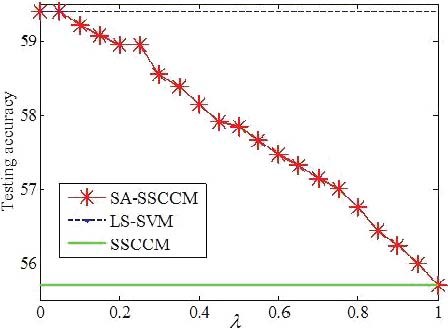
*j* =1

= ,

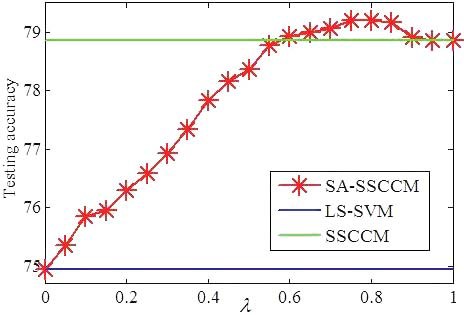
任何QP求解器都可以有效地解决这个问题。

在SA-SSCCM中简单地设置为SSCCM之后的全零，然后SA-SSCCM实际上仅从标记的实例开始。

在未标记数据损害分类的情况下，我们首先考虑bupa数据集。 SA-SSCCM中的正则化参数λ控制半监督和监督学习之间的权衡。 图1（a）显示了SA-SSCCM相对于[0,0.05,0.1，...，1]的不同λ值的空间为0.05的性能。 在图1（a）中，SSCCM的性能比LS-SVM差，表明在这种情况下未标记的数据会损害性能。 当λ接近零和一时，SA-SSCCM的性能退化为LS-SVM和SSCCM的性能，并随着λ的增加而减小。 当λ经过良好调整时，S3CCM的性能可以达到SSCCM和LS-SVM之间更好的性能，即LS-SVM的性能。 因此，在未标记数据损害分类的情况下，具有良好调整的λ的SA-SSCCM可以与LS-SVM相当。



(a)



(b)

图1. LS-SVM，SSCCM和SA-SSCCM的精度，λ在[0,0.05,0.1，...，1]范围内变化，超过（a）bupa和（b）电离层。

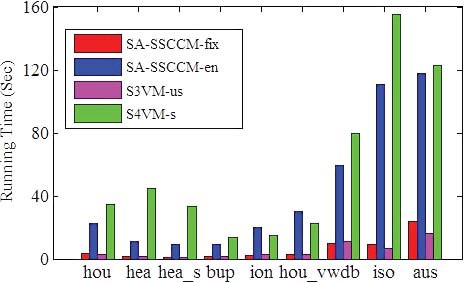
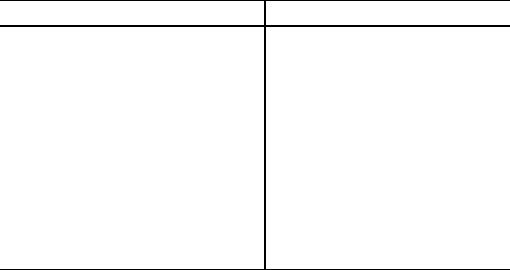


图2. SA-SSCCM\_fix，SA-SSCCM\_en，S3VM\_us和S4VM\_s的运行时间。

此外，在未标记数据有益于分类的情况下，我们对电离层数据集进行说明。 图1（b）显示了SA-SSCCM相对于电离层的[0,0.05,0.1，...，1]的不同λ值的性能。 在图1（b）中，SSCCM比LS-SVM表现更好，表明在这种情况下未标记的数据可以使分类受益。 SA-SSCCM的性能随着λ的增加而增加，并且在λ调整良好时优于SSCCM和LS-SVM。 因此，当未标记数据有利于分类时，具有良好调整的λ的SA-SSCCM可以胜过SSCCM和LS-SVM。

从这样的例子中，我们可以观察到SA-SSCCM的预测确实倾向于那些之间的平衡

表二

真实数据集的属性

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 数据 | #Ins | #有限元分析 | ID | 数据 | #所 | #有限元分析 |
| 1 | 屋 | 232 | 16 | 11 | 光明女神 | 690 | 15 |
| 2 | 心 | 270 | 9 | 12 | 糖尿病 | 768 | 8 |
| 3 | 心脏，statlog | 270 | 13 | 13 | 德语 | 1000 | 20 |
| 4 | 保柏 | 345 | 6 | 14 | Optdigits | 1143 | 42 |
| 5 | 电离层 | 351 | 34 | 15 | BCI | 400 | 117 |
| 6 | 车辆 | 435 | 16 | 16 | G241c | 1500 | 241 |
| 7 | 众议院投票 | 435 | 16 | 17 | G241d | 1500 | 241 |
| 8 | 心律失常 | 452 | 279 | 18 | COIL2 | 1500 | 241 |
| 9 | WDBC | 569 | 14 | 19 | Digit1 | 1500 | 241 |
| 10 | Isolet | 600 | 51 | 20 | 美国 邮政 | 1500 | 241 |

SSCCM和LS-SVM，它们完全验证了我们在第III-A节中的分析。 此外，具有良好调谐的λ的SA-SSCCM可以预期与LS-SVM和SSCCM之间的更好的相当，或甚至优于LS-SVM和SSCCM。 当标记数据足够时，我们可以通过交叉验证选择λ的值，如在监督场景中，当标记数据是恐慌时，我们转到[22]中的集合策略。 然而，正则化参数的选择对于无监督和半监督分类仍然是一个开放的问题，并且λ的值在未来仍然是值得研究的问题。

1. *绩效评估*

在这里，我们通过与监督的LS-SVM，半监督SSCCM和现成的安全半监督分类方法S3VM\_us [18]和S4VM [17]进行比较，评估SA-SSCCM在一组真实数据集上的性能。 对于SSCCM，我们采用（3）中的公式进行更直接的比较，对于S4VM，我们采用具有代表性采样的版本以实现更高效的计算[17]。 包括SA-SSCCM的电感半监督方法也可以通过决策函数预测给定的未标记数据。 因此，在我们的实验中，为了与转导S3VM\_us和S4VM进行比较，我们以转换方式实现所有方法，即学习标记和未标记数据，并预测给定未标记数据的性能。

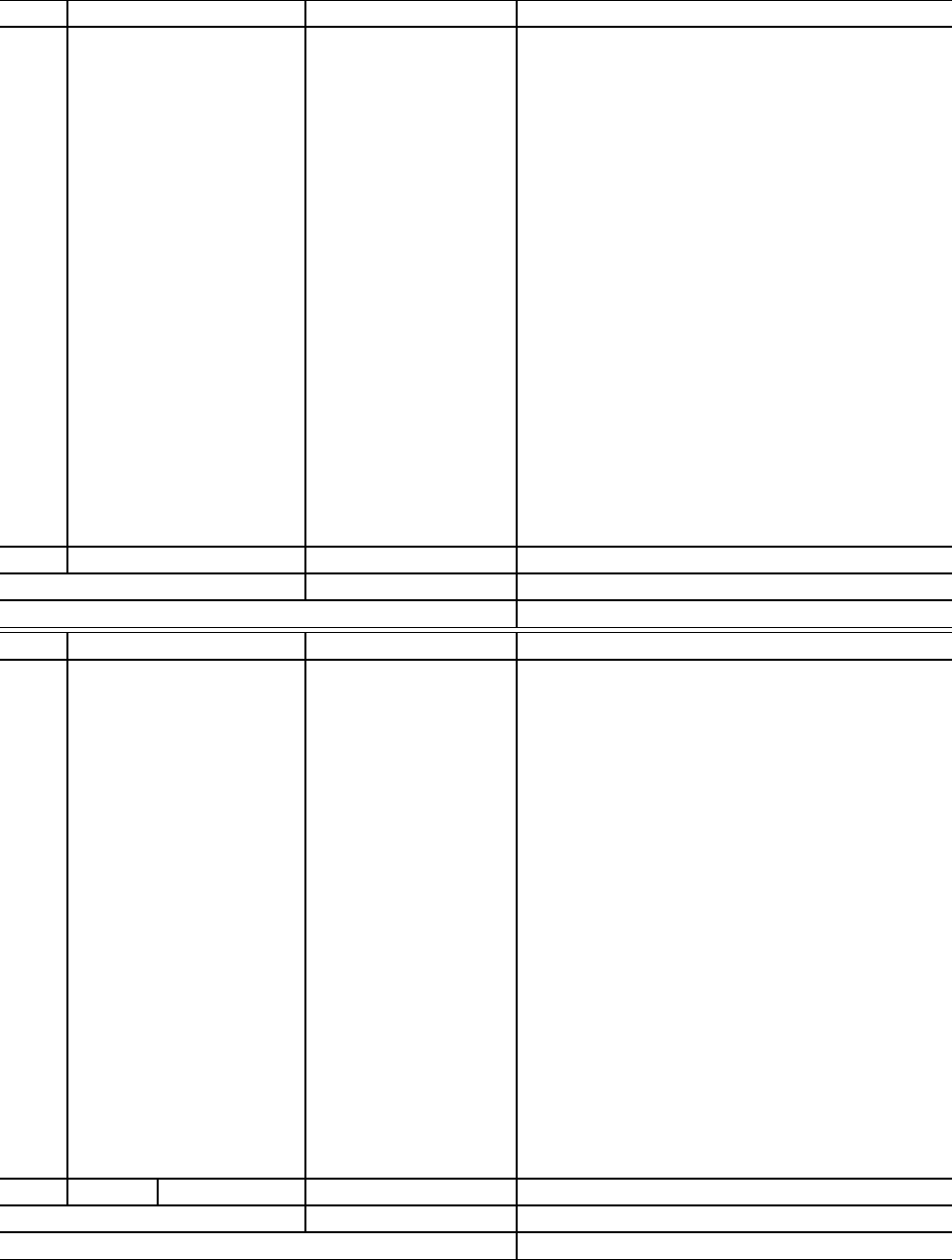
* 1. *实*验设置：我们对UCI和基准数据集进行性能比较，其描述见表II。 每个UCI数据集被随机分成两半，一半用于训练，另一半用于测试，训练集仅包含十个标记实例，并且休息符未标记。 该过程连同分类器学习重复30次，并报告平均测试精度。 这里采用线性和高斯核。 正则化参数C1和C2分别固定为1和0.1，λ1，λ2,和η是

固定为100,1和1，ε设定为10−3，并且

高斯核中的宽度参数σ设置为所有实例对之间的平均距离。

对于基准数据集，我们遵循[4]和[19]中的实验设置。 具体而言，对于每个数据集，都有

性能与十个标签的实例比较

sa-SSCCM\_fix SSCCM\_en

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 林。 | 支持向量机 | 最小二乘支持向量机 | TSVM | SSCCM | S3VM\_*us* | S4VM\_*s* |
| 1 | 92.0 ± 1.4 | 89.1 ± 1.2 | 93.6 ± 1.1*W* | 95.3 ± 3.9*W* | 92.3 ± 1.5*L* | 93.3 ± 1.8*L* |

91.1 ± 3.3*W* **95.2** ± **3.5***W*

*L*

2 58.6 ± 11.2 58.4 ± 7.8 58.2 ± 10.3 57.3 ± 9.7*L* 58.4 ± 10.5 58.8 ± 11.4 57.1 ± 8.7*L* **59.1** ± **9.8***W*

*W*

3 63.6 ± 8.0 65.7 ± 7.3 63.9 ± 7.5 **68.2** ± **11.6***W* 63.8 ± 8.2 63.9 ± 7.8 67.8±10.3*W*68.0±10.2*W*

4 56.6 ± 7.7 **59.4** ± **6.2** 55.3±7.2*L*55.7±5.3*L* 55.8 ± 7.6 55.8 ± 7.8 57.9 ± 5.6*L* 59.2 ± 5.9*W*

*W*

5 76.9 ± 5.6 74.9 ± 4.2 74.9±2.5*L*78.9±4.6*W*76.9±5.6W 76.6±4.5W 78.4±4.1*W* **78.9** ± **4.5***W*

6 79.1 ± 3.2 77.3 ± 3.1 76.0±5.9*L*81.6±7.6*W*78.8±5.4W 79.5±5.2W 80.1±5.8*W* 81.1 ± 7.2*W*

*L*

7 66.5 ± 11.7 64.1 ± 14.2 **7**7.6±10.9*W*74.2±13.7*W*68.2±11.8*W*70.6±9.2*W*74.5±7.6*W* 74.5 ± 7.7*W*

*L*

*L*

8 62.5 ± 13.8 58.0 ± 16.4 64.6±13.4*W*66.5±15.2*W*62.9±14.3L 63.2±13.2L 64.6±6.8*W* 64.8 ± 6.7*W*

*L*

*L*

9 84.3 ± 2.9 84.9 ± 2.4 83.5 ± 1.7 86.6 ± 2.7*W* 84.3 ± 2.9 84.8 ± 2.6*W* 84.5 ± 3.6*L* **86.7** ± **2.9***W*

10 92.6 ± 6.8 92.6 ± 5.5 87.9±9.4*L*91.4±2.7*L*93.4±7.4W 94.6±6.8*W* 92.6 ± 6.6*W* 94.0 ± 8.7*W*

*W*

*W*

11 76.7 ± 7.1 76.3 ± 12.0 80.8±9.4*W*83.2±5.6*W*79.4±7.8*W*79.8±7.6*W*81.3±5.1*W* 82.2 ± 5.3*W*

*L*

*L*

*L*

*L*

12 58.8 ± 2.8 59.4 ± 6.9 57.7±1.0*L*58.2±2.6*L*58.8±2.8W58.2 ± 3.4 59.2 ± 2.1*W* **59.6** ± **2.6***W*

13 60.1 ± 6.0 61.5 ± 4.8 56.2±9.2*L*62.6±6.9*W*61.6±6.2*W*60.2±5.6W60.8 ± 5.4*L* 62.2 ±6.8

*W*

14 95.9 ± 2.3 97.4 ± 1.3 99.7±0.4*W*99.6±0.2*W*96.4±2.6L 98.8±2.8*W*98.7±0.9*W* 98.9 ± 0.7*W*

15 52.8 ± 1.9 52.6 ± 3.1 51.3±3.4*L*53.6±2.6*W*52.6±2.3W52.0 ± 2.3 **55.6** ± **2.4***W* 54.8 ± 3.0*W*

*W*

*W*

16 54.6 ± 4.4 61.3 ± 3.5 **7**8.4±5.8*W*63.2±3.4*W*54.7±2.9L 54.6±4.2L62.1 ± 3.5*L* 63.8 ± 3.6*W*

17 56.4 ± 5.2 59.5 ± 4.1 53.4±9.6*L*55.4±7.4*L*56.4±5.4W 56.3±5.1W 59.0±3.5W**59.7** ± **4.4***W*

18 56.3 ± 5.1 58.5 ± 4.3 56.5 ± 5.9 **58.9** ± **4.9** 56.3 ± 5.9 56.4 ± 4.8 58.5 ± 4.4 58.8 ± 4.9

19 77.0 ± 5.2 65.4 ± 10.8 **79.7** ± **3.3***W* 65.6 ± 9.1 77.4 ± 5.4*L* 76.9 ± 5.1*L* 65.6 ± 4.6 66.5 ± 8.4

20 77.9 ± 3.0 79.9 ± 0.9 72.4 ± 3.7*L* 80.1 ± 0.4 78.0 ± 2.7*W* 78.2 ± 2.9*W* 80.2 ± 0.9 **80.3** ± **0.7**

平均。 ACC 69.9 69.8 71.1 71.8 70.3 70.6 71.5 **72.4**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Semisuperv。 与superv。 W / T / L | | | 7/4/9 | 12/3/5 | 3/17/**0** | 4/16/**0** | 9/9/2 | 12/8/**0** |
| 安全的semisuperv。 与semisuperv。 W / T / L | | |  |  | 8/5/7 | 7/7/6 | 5/9/8 | 6/12/**2** |
| 高铭。 | 支持向量机 | 最小二乘支持向量机 | TSVM | SSCCM | S3VM\_*us* | S4VM\_*s* sa-SSCCM\_fix SSCCM\_en | | |
| 1 | 91.0 ± 1.1 | 89.6 ± 0.6 | 92.8 ± 1.0*W* | **95.6** ± **1.6***W* | 91.5 ± 1.0*L* | 92.4±1.1*W*92.8±0.9*W* 95.3 ± 0.8*W*  *L* | | |

2 57.2 ± 8.6 57.4 ± 7.2 58.8 ± 7.9*W* 58.1 ± 6.4 5**6.7±8.5L 59.8±7.8***W* 57.6 ± 6.4 58.1 ± 6.4

3 61.9 ± 3.9 61.8 ± 5.1 61.4 ± 4.8 64.1 ± 4.9*W* 62.1 ± 3.9 62.2 ± 4.1 **64.9** ± **4.8***W* 64.1 ± 4.9*W*

4 **61.9** ± **6.3** 60.9 ± 2.4 59.1 ± 5.9*L* 60.6 ± 6.4 59.4±5.4*L*61.6±4.6W61.4 ± 5.7 60.9 ± 6.4

5 77.9 ± 4.9 77.8 ± 6.9 73.7±3.3*L*79.3±5.0*W*77.8±4.9W 77.4±4.5W 79.7±4.7*W* **79.3** ± **5.0***W*

6 78.1 ± 5.4 77.8 ± 3.7 73.5±7.0*L*77.2±6.1*L*77.5±6.1W 78.8±5.6W77.8 ± 6.1 **80.5** ± **4.8***W*

*W*

7 71.4 ± 12.4 69.6 ± 9.8 74.9±13.5*W*76.7±8.7*W*70.5±11.6L 72.6±10.8L 74.3±5.0*W*

*L*

8 73.4 ± 11.9 72.6 ± 10.2 **7**8.8±7.5*W*75.0±9.3*W*72.6±11.0L 74.5±11.6L 76.2±9.8*W*

*W*

76.3 ± 4.7*W*

75.2 ± 10.2*W*

9 **85.4** ± **2.3** 82.7 ± 3.9 84.2 ± 1.5 82.4 ± 5.0 85.3 ± 2.3*W* 84.8 ± 1.8 82.0 ± 4.1 82.5 ± 4.9

10 89.1 ± 7.3 88.3 ± 8.4 87.8 ± 9.4 87.0±7.1*L*89.8±7.3W88.6 ± 6.8 **89.4** ± **6.9***W* 87.1 ± 14.6

*W*

11 77.1 ± 5.9 78.0 ± 6.4 81.9±7.3*W*83.0±5.1*W*78.7±7.6*W*77.9±6.2L 79.2±5.2*W* **83.2** ± **4.8***W*

*L*

*L*

12 59.1 ± 3.1 59.8 ± 5.2 51.7 ± 1.0*L* 59.6 ± 2.6 58.8 ± 2.8*W* 57.6 ± 2.6*W* 59.2 ± 2.1 **59.7** ± **2.6**

13 60.6 ± 5.6 59.6 ± 8.8 53.3±6.3*L*62.8±6.9*W*61.1±5.1W 60.8±5.4W 61.2±5.3*W* 62.5 ± 6.9*W*

*L*

14 91.6 ± 3.8 92.6 ± 2.3 96.8±1.6*W*97.2±1.4*W*92.3±3.1L 94.8±3.1*W*97.7±1.1*W* **98.4** ± **1.5***W*

*L*

*W*

15 51.1 ± 2.9 50.9 ± 0.8 51.4 ± 2.7 53.0 ± 2.4*W* 51.1 ± 2.7 51.2 ± 2.7 **55.1** ± **3.6***W*

*W*

16 53.6 ± 4.4 52.4 ± 4.6 **5**9.0±4.8*W*49.9±0.1*L*53.6±4.6L 53.8±4.4L 55.8±2.2*W*

*W*

17 52.8 ± 5.2 52.2 ± 4.6 53.1 ± 6.7 49.9 ± 0.2*L* 52.9 ± 5.0 53.0 ± 5.4 53.2 ± 2.0*W*

*W*

18 56.8 ± 4.0 53.0 ± 3.0 57.3 ± 4.2 55.7 ± 2.9*W* 56.8 ± 4.1 57.0 ± 3.8 **60.7** ± **3.8***W*

*W*

53.0 ± 2.4*W*

57.5 ± 2.8*W*

*W*

**58.3** ± **3.0***W*

*W*

55.8 ± 2.8*W*

19 57.5 ± 6.5 56.1 ± 12.1 **8**0.3±3.1*W*66.3±10.7*W*60.2±6.7*W*76.8±6.4*W*60.2±2.5*W*66.6±10.6*W*

*L*

*L*

*L*

20 79.9 ± 2.2 80.0 ± 0.1 71.3±2.6*L*81.1±0.9*W*79.5±1.9W 79.3±2.3W 81.9±1.5*W* 81.4 ± 1.1

平均。 acc 69.1 68.4

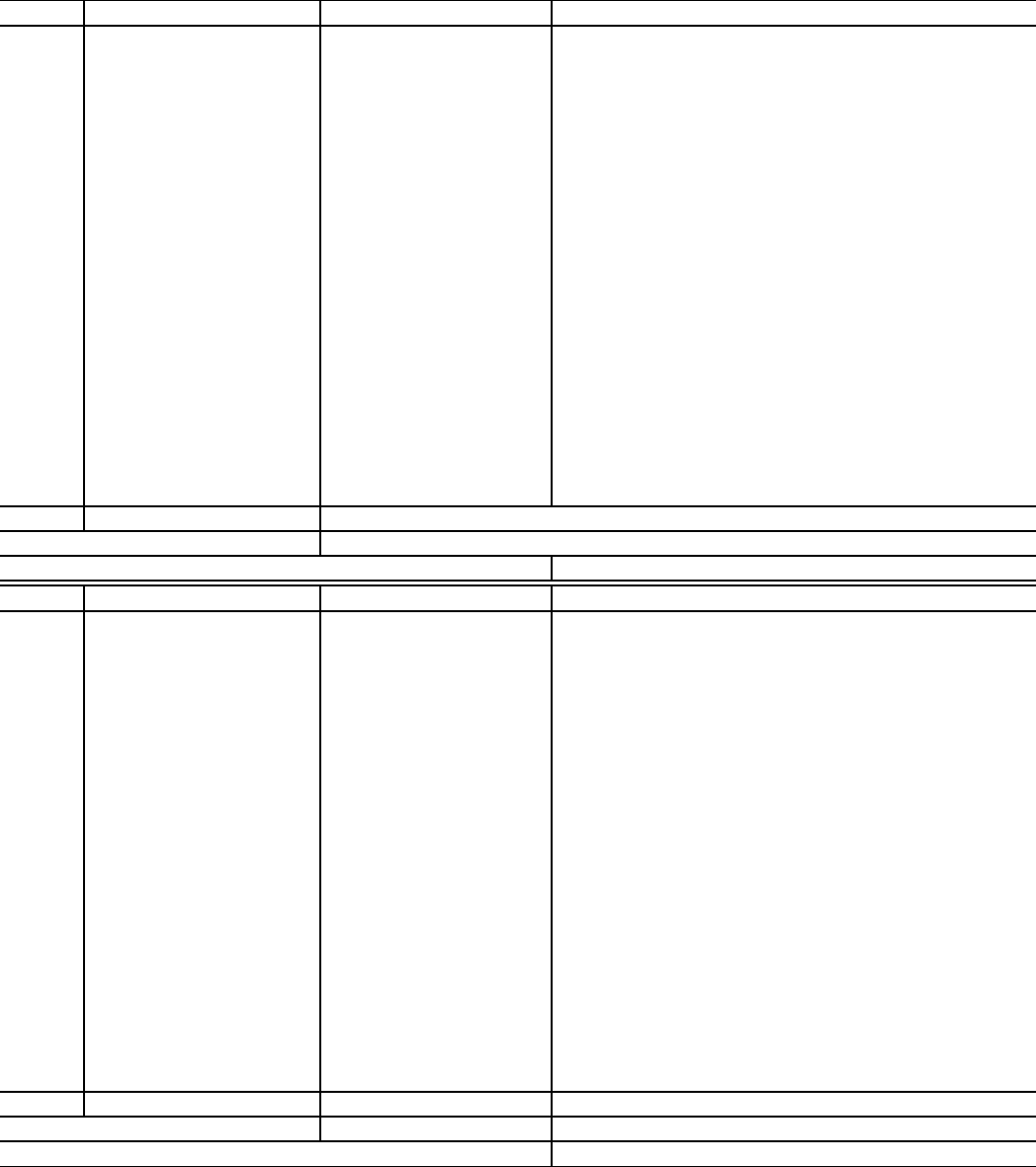
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 70.1 | 70.7 | 69.4 | 70.7 | 71.0 | **71.8** |
| 7/7/6 | 11/5/4 | 2/17/1 | 4/18/**0** | 15/7/**0** | 14/6/**0** |
|  |  | 7/5/8 | 7/9/6 | 6/9/5 | 4/16/**0** |

Semisuperv。 与superv。 （W / T / L）

安全的semisuperv。 与semisuperv。 （W / T / L）

两个设置，一个包括十个标记实例，另一个包括100个实例。 此外，对于每个数据集和每个设置，有12个标记数据子集，最后是

报告未标记数据的平均表现。 比较方法中的正则化参数C1和C2分别设置为100和0.1，λ1，λ2,和η是固定的

性能与100个标签实例的比较

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 支持向量机 | 最小二乘支持向量机 | TSVM | SSCCM | S3VM\_*us* | S4VM\_*s* | SA-SSCCM\_*cv* | SA-SSCCM\_*en* |
| 1 | 96.7 ± 0.4 | 96.7 ± 0.4 | 96.6 ± 0.5 | 96.7 ± 0.4 | 95.6 ± 0.5*L* | 96.5 ± 0.4 | 96.7 ± 0.4 | **96.9** ± **0.4** |

2 85.2 ± 1.9 85.8 ± 2.0 85.1 ± 1.9 **86.2** ± **2.0** 85.1 ± 1.8 85.5 ± 1.6 86.1 ± 2.1 86.2 ± 2.1

3 82.3 ± 2.2 83.8 ± 1.5 81.9 ± 1.3 84.1 ± 1.5 82.9 ± 2.1 82.0 ± 1.8 84.5 ± 1.7 **84.7** ± **1.6**

4 59.5 ± 3.7 58.5 ± 2.8 57.7 ± 3.6*L* 59.2 ± 3.0 59.1 ± 3.7*W* 59.3 ± 3.1*W* **59.7** ± **2.9** 58.4 ± 2.8

5 89.6 ± 0.9 88.9 ± 0.9 86.7 ± 3.1*L* 89.6 ± 0.6*W* 89.6 ± 0.9*W* **90.6** ± **0.8***W* 90.1 ± 0.5*W* 89.5 ± 0.7*W*

6 92.4 ± 1.9 90.9 ± 2.1 88.5 ± 1.6*L* 91.8 ± 1.9*W* 92.4 ± 1.9*W* **92.8** ± **2.1***W* 92.7 ± 2.1*W* 91.8 ± 1.9*W*

*W*

7 91.6 ± 1.8 91.4 ± 1.9 87.3 ± 2.6*L* 93.6 ± 1.9*W* 91.7 ± 2.7*W* 91.4 ± 1.5*W* **93.9** ± **1.9***W* 93.3 ± 2.1*W*

8 91.8 ± 1.8 91.9 ± 18 90.3 ± 2.4*L* 94.0 ± 1.3*W* 91.8 ± 2.2*W* 92.2 ± 2.1*W* **94.1** ± **1.3***W* 93.9 ± 1.4*W*

9 95.6 ± 0.9 95.7 ± 0.6 83.6 ± 0.9*L* 96.0 ± 0.5*W* 95.6 ± 0.9*W* 94.9 ± 0.8*W* **96.8** ± **0.5***W*

*W*

10 98.0 ± 1.2 97.5 ± 1.0 99.6 ± 0.3*W* 99.9 ± 0.1*W* 99.1 ± 1.3*W* 98.2 ± 0.9*L* **100** ± **0***W*

*W*

96.0 ± 0.5*W*

**100** ± **0***W*

*W*

11 96.4 ± 4.6 95.8 ± 4.3 97.6 ± 5.1*W* **98.0** ± **5.1***W* 97.5 ± 4.6*W* 96.7 ± 3.9 **98.0** ± **5.0***W* 97.7 ± 5.1*W*

12 73.4 ± 2.2 73.8 ± 1.3 70.6 ± 6.9*L* 76.1 ± 1.6*W* 73.6 ± 2.1*W* 74.8 ± 1.8*W* 76.8 ± 0.9*W* **77.1** ± **0.9***W*

*W*

*W*

13 69.9 ± 2.2 63.5 ± 1.9 58.2 ± 18.7*L* 72.2 ± 0.1*W* 69.5 ± 1.7*W* 70.1 ± 1.5*W* **72.2** ± **0.1***W* 70.3 ± 0.5*W*

*L*

14 99.1 ± 0.4 98.9 ± 0.5 **99.8** ± **0.2 99.8**±**0***W* 99.4 ± 0.4 99.6 ± 0.2 **99.8** ± **0***W* 99.6 ± 0.3*W*

15 71.8 ± 3.5 73.1 ± 2.9 71.5 ± 4.1 75.3 ± 2.4*W* 71.6 ± 3.2 71.5 ± 3.4 76.2 ± 2.2*W* **76.8** ± **2.9***W*

*W*

*W*

16 75.0 ± 1.9 76.3 ± 1.5 **80.0** ± **1.7***W* 73.9 ± 2.2*L* 75.0 ± 1.9*L* 75.2 ± 2.5*L* 76.5 ± 1.5*W* 75.6 ± 1.8*W*

17 72.4 ± 2.8 73.1 ± 3.1 **76.3** ± **2.4***W* 73.8 ± 2.7 72.8 ± 1.8*L* 74.9 ± 2.3*W* 74.1 ± 2.5 72.9 ± 2.1

*L*

18 79.6 ± 2.1 80.2 ± 2.5 80.4 ± 1.7 80.8 ± 2.2 79.2 ± 2.3 **81.1** ± **2.9***W* 80.8 ± 3.1 80.0 ± 2.7

19 92.4 ± 1.4 92.3 ± 1.5 92.5 ± 1.9 91.6 ± 1.8*L* 92.4 ± 1.4 **92.6** ± **2.1** 92.3 ± 1.5*W* 92.5 ± 1.5*W*

20 88.2 ± 0.9 88.0 ± 0.8 86.7 ± 1.4*L* 86.6 ± 1.1*L* 88.1 ± 0.9*W* **88.5** ± **1.3***W* 88.2 ± 0.8*W* 87.8 ± 1.1*W*

平均。 ACC 85.0 84.8 83.5 85.9 85.1 85.4 **86.5** 86.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Semisuperv。 与superv。 （W / T / L） | 4/7/9 | 11/6/3 | 2/17/1 | 3/17/**0** | 11/9/**0** | 11/9/**0** |
| 安全的semisuperv。 与semisuperv。 （W / T / L） |  |  | 9/9/2 | 9/8/3 | 8/12/**0** | 5/14/1 |
| 数据集 支持向量机 最小二乘支持向量机 | TSVM | SSCCM | S3VM\_*us* | S4VM\_*s* | SA-SSCCM\_*cv* | SA-SSCCM\_*en* |
| 1 96.7 ± 0.4 96.7 ± 0 | 96.6 ± 0.4 | 96.7 ± 0 | 96.4 ± 0.5 | 96.6 ± 0.4 | **97.0** ± **0** | **97.0** ± **0** |

2 86.0 ± 1.4 87.8 ± 0.5 81.4 ± 1.8*L* 87.4 ± 0.7*L* 85.7 ± 1.8*W* 85.9 ± 1.2*W* **87.9** ± **0.7***W* 87.6 ± 0.7

3 79.6 ± 0 84.7 ± 0 80.0 ± 0*W* 86.5 ± 0*W* 79.4 ± 0*L* 80.5 ± 0*W* 86.5 ± 0*W* **86.9** ± **0***W*

*L*

*W*

4 66.3 ± 2.9 65.3 ± 3.4 66.5 ± 2.7 65.7 ± 3.2 66.0 ± 3.7 **66.7** ± **3.5** 66.5 ± 3.7*W* 65.9 ± 3.2

*W*

5 91.5 ± 1.7 91.9 ± 0.9 81.6 ± 2.0*L* 92.2 ± 0.7*W* 89.6 ± 0.9*L* 90.7 ± 0.8*W* **92.3** ± **0.6***W* 92.2 ± 0.8*W*

*W*

6 92.2 ± 3.4 95.0 ± 1.1 92.0 ± 1.2 **96.1** ± **1.0***W* 92.4 ± 1.1 94.3 ± 0.9*W* **96.1** ± **1.0***W* 96.0 ± 1.0*W*

*W*

7 92.4 ± 1.7 94.1 ± 2.3 88.2 ± 3.2*L* **94.7** ± **1.5** 92.5 ± 2.3*W* 91.9 ± 1.9*W* **94.7** ± **2.2** 94.5 ± 2.3

8 89.6 ± 2.6 92.4 ± 3.3 85.6 ± 2.6*L* 94.5 ± 2.8*W* 89.8 ± 3.1*W* 89.3 ± 3.4*W* 94.5 ± 2.2*W* 94.5 ± 2.8*W*

9 94.9 ± 0.7 95.1 ± 0.4 89.8 ± 1.6*L* 96.9 ± 0.9*W* 94.9 ± 1.1*W* 94.8 ± 0.9*W* 97.1 ± 0.4*W* 96.5 ± 0.8*W*

10 97.2 ± 1.2 98.6 ± 0.8 99.8 ± 0.2*W* **100** ± **0***W* 98.4 ± 0.6*W* 99.8 ± 0.9*W* **100** ± **0***W* **100** ± **0***W*

11 95.3 ± 0.2 95.8 ± 0.4 95.0 ± 0.2 96.0 ± 0.4 95.5 ± 0.2 95.3 ± 0.2 **96.3** ± **0.5***W* 95.9 ± 0.6

12 73.2 ± 1.8 75.1 ± 1.0 71.2 ± 1.3*L* 74.2 ± 1.0*L* 72.6 ± 1.0*W* 73.4 ± 1.0*W* **75.4** ± **1.7***W* 75.2 ± 1.9*W*

13 71.6 ± 3.4 67.6 ± 2.9 53.1 ± 3.2*L* 72.7 ± 3.7*W* 71.0 ± 2.3*W* 71.4 ± 2.9*W* 72.9 ± 2.4*W* 71.8 ± 3.1*W*

14 99.6 ± 0.1 99.7 ± 0.2 99.7 ± 0.1*W* 99.7 ± 0.1 99.4 ± 0.2*L* 99.6 ± 0.1*L* **99.8** ± **0***W* 99.7 ± 0.1*W*

*L*

*W*

15 67.4 ± 3.2 68.6 ± 2.3 65.8 ± 2.8*L* 70.2 ± 3.1*W* 67.2 ± 3.4*W* 67.2 ± 2.8*W* **70.8** ± **2.6***W* 69.8 ± 3.8*W*

16 69.6 ± 6.8 68.4 ± 5.1 **78.1** ± **2.2***W* 77.3 ± 4.1*W* 69.8 ± 3.2*L* 75.2 ± 2.9*W* 77.9 ± 2.1*W* 75.2 ± 3.3*W*

*L*

*L*

17 61.3 ± 8.6 64.2 ± 7.8 **66.3** ± **5.8***W* 62.8 ± 6.9*L* 61.9 ± 6.2*L* 62.3 ± 3.3*W* 65.8 ± 4.2*W* 64.1 ± 3.6*W*

*L*

*W*

18 87.6 ± 2.8 87.2 ± 2.5 87.5 ± 2.5 88.3 ± 2.2*W* 87.5 ± 2.9 87.2 ± 3.1 88.8 ± 2.1*W* **89.2** ± **3.0***W*

*W*

19 95.1 ± 1.6 94.8 ± 1.8 95.5 ± 1.5 95.5 ± 1.6 95.1 ± 1.4 **96.2** ± **1.3***W* 95.8 ± 0.9*W* 95.2 ± 1.8

20 84.6 ± 2.1 85.7 ± 1.8 **90.8** ± **1.5***W* 87.9 ± 2.6*W* 86.8 ± 2.8*W* 89.4 ± 1.5*W* 88.2 ± 1.5*W* 87.8 ± 2.1*W*

*L*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 平均。 ACC 84.6 85.3 | 83.2 | 86.7 | 84.3 | 85.4 | **87.2** | 86.8 |
| Semisuperv。 与superv。 （W / T / L） | 6/6/8 | 11/6/3 | 2/15/3 | 7/13/**0** | 16/4/**0** | 12/8/**0** |
| 安全的semisuperv。 与semisuperv。 （W / T / L） |  |  | 8/7/5 | 10/7/3 | 3/17/**0** | 3/16/1 |

分别为100,0.1和1，并且ε设定为10−3。 使用线性和高斯核，高斯核中的宽度参数σ被设置为当标记10个实例时由δ表示的实例之间的平均距离，并且通过来自{0.25,0.5,1,2的5倍交叉验证来选择，在100个标记实例的情况下，标记的训练数据上的4}δ。

在十个标记实例的情况下，我们分别通过将其固定为0.5和整体策略（分别对应于SA-SSCCM\_fix和SA-SSCCM\_en）来设置SA-SSCCM中的λ值。 在100个标记实例的情况下，我们分别通过集合和交叉验证策略设置其值，分别对应于SA-SSCCM\_en和SA-SSCCM\_cv。 另外，在SA-SSCCM\_en中，

λ的候选值从区间[0,1]中均匀选择，间隔为0.2，即[0,0.2,0.4,0.6,0.8,1]。

* 1. *性*能比较：具有10和100个标记实例的性能分别示于表III和IV中，其中上半部分和下半部分分别对应于采用线性和高斯核的性能。 在每个部分中，每行（除了最后三行）给出了每个数据集上各个方法的性能（分类准确度和标准偏差）。 在每一行中，粗体值表示每个数据集的最佳性能，带有上标W / L的值表示半监督或安全的半监督方法比相应的监督对应物表现更好/更差，通过t检验具有显着差异并且下标W / L的值表示安全半监督方法比原始半监督方法表现更好/更差，并且通过t检验具有显着差异。 倒数第三行分别提供了所有数据集中各个方法的平均精度。 倒数第二行给出了半监督或安全半监督方法比相应的监督对应方执行更好/可比/更差（W / T / L）的情况数，最后一行给出了数字安全半监督方法比原始半监督方法表现更好/可比/更差（W / T / L）的情况。

从这些表中，我们可以做出如下几点观察。

1. SVM和LS-SVM的整体性能具有可比性。 然而，SSCCM（基于LS-SVM）的整体性能优于TSVM（基于SVM），因为修改的聚类假设可以比聚类假设更好地捕获实际数据分布。 这正是我们开发基于SSCCM的安全感知半监督分类方法的原因，尽管这种机制也可以类似地应用于其他半监督分类方法，例如TSVM。
2. S3VM\_us中性能退化的可能性远小于TSVM。 具体来说，TSVM在80个案例中的32个中完全比SVM差，而S3VM\_us仅在5个案例中表现比SVM差。 同时，S4VM从未比SVM表现更差，其整体性能与TSVM相比具有很强的竞争力。 因此，可以采用S3VM\_us和S4VM进行安全的半监督分类。
3. 在15例情况下，SSCCM的表现比LS-SVM差，而SA-SSCCM在所有情况下都比LS-SVM表现更好。 因此，SA-SSCCM可用于安全的半监督分类。
4. 预计安全半监督方法的表现不会比监督对应方差，并且同时表现不比原始的半监督方法差。 在22例情况下，S3VM\_us的性能比TSVM差，而S4VM则执行

18例中完全比TSVM差，而SA-SSCCM仅在2例中表现比SSCCM差。 因此，可以期望SA-SSCCM实现安全半监督分类的新目标。

1. SA-SSCCM的整体性能优于S3VM\_us和S4VM，表明SA-SSCCM在安全半监督分类方面具有很高的竞争力。
   1. *计*算效率：为了说明SA-SSCCM的计算效率，我们提供了SA-SSCCM与现成的安全半监督分类方法S3VM\_us和S4VM的运行时间比较。 对于S4VM，我们基于代表性采样选择S4VM\_s，因为它比基于全局模拟退火搜索的S4VM\_a更快地学习[17]。

图2显示了SA-SSCM\_fix，SA-SSCCM\_en，S3VM\_us和S4VM的训练时间，在10个标记实例的情况下，线性内核超过9个代表性数据集。 从图2可以看出，SA-SSCCM\_fix具有与S3VM\_us相当的时间成本，并且比S4VM\_s具有更低的时间成本。 SA-SSCCM\_en比S3VM\_us更差，但在大多数情况下优于S4VM\_s。 因此，与现成的安全半监督分类方法相比，SA-SSCCM的效率具有竞争力。

1. 结论

由于在某些情况下使用未标记的数据，半监督分类方法可能比其受监督的对应方法产生更差的性能。 因此，它自然地降低了将半监督分类方法应用于实际应用的可信度。 为了在不牺牲半监督分类性能的情况下解决这个问题，我们通过根据可用的未标记数据自适应地控制半监督和监督学习之间的权衡来发明安全控制机制。 在实施过程中，我们根据最近的SSCCM开发了一种安全感知的半监督分类方法，称为安全感知SSCCM。 SA-SSCCM的预测在未标记数据方面完成了SSCCM和LS-SVM之间的权衡。 最后，几个真实数据集的实证结果表明，SA-SSCCM的整体性能优于SSCCM和LS-SVM，此外，SA-SSCCM的性能从未明显低于LS-SVM，并且很少明显逊于SSCCM。 同时，与现成的安全半监督分类方法相比，SA-SSCCM的计算效率具有竞争力。

参数λ的值是SA-SSCCM中控制SSCCM和LS-SVM之间权衡的关键问题。 在本文中，可以在标记数据足够时通过交叉验证来选择，否则通过集合策略选择。 然而，在半监督学习中寻求参数的最优值仍然是一个值得研究的开放性问题，进一步建立一个与之相关的可靠理论是绝对必要的，它自然成为我们未来的重要工作。

参考

1. Z.-H. Zhou和M. Li，“通过分歧进行半监督学习”

*K*nowl。 天道酬勤。 Syst。，vol。 24，不。 3，pp.415-439,2010。

1. X. Zhu和AB Goldberg，半监督学习导论。 美国加利福尼亚州圣马特奥市：Morgan Kaufmann，2009年。
2. X. Zhu，“半监督学习文学调查”，博士。 论文，部门。 Sci。，Wisconsin-Madison，Univ。，Madison，WI，USA，Jul.2008。
3. O. Chapelle，B。Scholkopf和A. Zien，半监督学习。 美国马萨诸塞州剑桥：麻省理工学院出版社，2006年。
4. PK Mallapragada，R。Jin，AK Jain和Y. Liu，“Semi-boost：Boosting for semi-supervised learning，”IEEE Trans。 模式肛门。 马赫。 Intell。，vol。 31，不。 11，pp.2000-2014，2009年11月。
5. T. Joachims，“使用支持向量机进行文本分类的转导推理”，Proc。 第16届国际 CONF。 马赫。 Learn。，1999，pp.200-209。
6. G. Fung和OL Mangasarian，“用于无标签数据分类的半监督支持向量机”，Optim。 方法Softw。，vol。 15，不。 1，pp.99-105,2001。
7. R. Collobert，F。Sinz，J。Weston和L. Bottou，“大规模转导SVM”，J。Mach。 学习。 Res。，vol。 7，pp.1687-1712，2006年1月。
8. Y.-F. Li，J。Kwok和Z.-H. 周，“使用标签均值的半监督学习”，Proc。 第26届国际 CONF。 马赫。 Learn。，2009，pp.633-640。
9. Y. Bengio，OB Alleau和NL Roux，“标签传播和二次标准”，在Semi-Supervised Learning，Cambridge，MA，USA：MIT Press，2006，pp.193-216。
10. X. Zhu和Z. Ghahramani，“学习带有标签传播的带标签和未标记数据”，Lang。 Technol。，Carnegie Mellon Univ。，Pittsburgh，PA，USA，Tech。 Rep.CMU-CALD-02-107,2002。
11. A. Blum和S. Chawla，“使用图形切口学习标记和未标记的数据”，在Proc。 第18届国际 CONF。 马赫。 Learn。，2001，pp.19-26。
12. M. Belkin，P。Niyogi和V. Sindhwani，“流形正则化：从标记和未标记的例子中学习的几何框架”，J。Mach。 学习。 Res。，vol。 7，不。 1，pp.2399-2434,2006。
13. B*. Pfahringer，“一个半监督的垃圾邮件检测器”，在Proc。 17th Eur。 CONF。 马赫。 学习。 10欧元 CONF。 原则实践。 Knowl。 Discovery Databases，2006，pp.1-5。*
14. N. Kasabov和S. Pang，“转导支持向量机和生物信息学中用于启动子识别的应用”，Neural Inf。 处理。 快报。 Rev.，vol。 3，第31-38页，2004年5月。
15. T. Yang和CE Priebe，“模型错误指定对半监督分类的影响”，IEEE Trans。 模式肛门。 马赫。 Intell。，vol。 33，不。 10，pp.2093-2103，2011年10月。
16. Y.-F. 李和Z.-H. 周，“将未标记的数据从未受到伤害”，Proc。 28th Int。 CONF。 马赫。 Learn。，2011，pp.1081-1088。
17. Y.-F. 李和Z.-H. 周，“通过无标签实例选择改进半监督支持向量机”，Proc。 第25届AAAI会议 ARTIF。 Intell。，2011，pp.500-505。
18. Y. Wang，S。Chen和Z.-H. 周，“基于修正聚类假设的新型半监督分类方法”，IEEE Trans。 神经网络。 学习。 Syst。，vol。 23，不。 5，pp.689-702，2012年5月。
19. B. Yang和S. Chen，“基于地点的无监督维度减少的伪装歧视”，Int。 J.模式识别。 ARTIF。 Intell。，vol。 24，不。 7，pp.1011-1025,2010。
20. J. Gorski和F. Pfeuffer，“Biconvex使用双凸函数进行设置和优化：调查和扩展，”数学。 方法操作。 Res。，vol。 66，不。 3，pp.373-407,2007。
21. Z. Xu，R。Jin，I。King，MR Lyu和Z. Yang，“适应性转换支持向量机的正则化”，Proc。 进阶 神经信息。 处理。 Syst。，2009，pp.2125-2133。
22. K. Chen和S. Wang，“正规化促进半监督学习”，Proc。 进阶 神经信息。 处理。 Syst。，2008，pp.281-288。
23. M. Gonen和E. Alpaydin，“多核学习算法”，J。Mach。 学习。 Res。，vol。 12，pp.2211-2268，2011年1月。

汪云云于2006年获得中国芜湖安徽师范大学计算机科学与技术学士学位，并获得博士学位。 2012年获得南京航空航天大学计算机科学与工程学士学位。

现任南京邮电大学计算机科学与工程系南京分院讲师。 她目前的研究兴趣包括模式识别，机器学习和神经计算。

陈松灿于1983年获得杭州大学（现浙江大学）数学学士学位，1983年获得上海交通大学计算机应用硕士学位，1985年获博士学位。 1997年通信和信息系统学位。

他于1986年在南京中国南京航空航天大学（NUAA）工作。自1998年以来，他一直是计算机科学与工程系的全职教授，

南京航空航天大学。 他撰写或合着了130多篇科学同行评审论文。 他目前的研究兴趣包括模式识别，机器学习和神经计算。