doi:10.3969/j.issn.1000-436x.2013.05.005

类不均衡的半监督高斯过程分类算法

夏战国，夏士雄，蔡世玉，万玲
（中国矿业大学 计算机科学与技术学院，江苏 徐州 221116）

摘 要：针对传统的监督学习方法难以解决真实数据集标记信息少、训练样本集中存在类不均衡的问题，提出了类不均衡的半监督高斯过程分类算法。算法引入自训练的半监督学习思想，结合高斯过程分类算法计算后验概率，向未标记数据中注入类标记以获得更多准确可信的标记数据，使得训练样本的类分布相对平衡，分类器自适应优化以获得较好的分类效果。实验结果表明，在类不均衡的训练样本及标记信息过少的情况下，该算法通过自训练分类器获得了有效标记，使分类精度得到了有效提高，为解决类不均衡数据分类提供了一个新的思路。
关键词：类不均衡；半监督；高斯过程分类；自训练
中图分类号：TP181 **文献标识码**：A **文章编号**：1000-436X(2013)05-0042-10

Semi-supervised Gaussian process classification algorithm addressing the class imbalance

XIA Zhan-guo, XIA Shi-xiong, CAI Shi-yu, WAN Ling
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: The traditional supervised learning is difficult to deal with real-world datasets with less labeled information when the training sets class is imbalanced. Therefore, a new semi-supervised Gaussian process classification of addressing was proposed. The semi-supervised Gaussian process was realized by calculating the posterior probability to obtain more accurate and credible labeled data, and embarking from self-training semi-supervised methods to add class label into the unlabeled data. The algorithm makes the distribution of training samples relatively balance, so the classifier can adaptively optimized to obtain better effect of classification. According to the experimental results, when the circumstances of training set are class imbalance and much lack of label information, The algorithm improves the accuracy by obtaining effective labeled in comparison with other related works and provides a new idea for addressing the class imbalance is demonstrated.
Key words: class imbalance; semi-supervised; Gaussian process classification; self-training

1 引言

高斯过程[1]是近年在国际上机器学习研究的热点领域之一。高斯过程是基于贝叶斯框架的无参数核方法，可用于有监督学习，被成功应用于回归与分类[2~5]。与人工神经网络（ANN）和支持向量机（SVM）相比，高斯过程的优点主要有 3 点：1) 在不牺牲性能的条件下容易实现，在模型构建过程中自动地获取超参数，具有完全的贝叶斯公式化表示，预测输出具有清晰的概率解释，并且可以直接实现多分类[6~13]；2) 由于高斯过程采用核函数，使它具有较强的非线性性能，可以解决线性不可分和特征维数过多的问题，从而在一定程度上避免了 ANN 中的“维数灾难”问题[14]；3) 高斯过程为贝叶斯学习提供了一个范式，根据训练样本可以从先验分布转换到后验分布，并可以对核函数的超参数进行推理；而 SVM 对超参数的选择却通常只能采用经验法或交叉验证方法[15,16]。因此，在数据挖掘、模式识别、

收稿日期：2012-05-29；**修回日期**：2013-02-07
基金项目：国家自然科学基金资助项目（50674086）；国家教育部博士点基金资助项目(20110095110010)
Foundation Items: The National Natural Science Foundation of China (50674086); The Ph.D. Programs Foundation of the Ministry of Education of China (20110095110010)

故障检测、生物医学、图像及文本识别等领域得到了广泛的应用，并已取得了较为显著的研究。

传统的监督学习方法只利用标记数据进行模型训练，一旦标记数据量很少，就会导致训练集不足，模型的泛化能力得不到保证，势必会引起回归或分类结果的精度下降，最终难以解决实际应用的问题。而在现实世界中，难以获得过多的已标记数据，对无标记数据进行标记则需要昂贵的代价，且易于标记错误。针对这种情况，学者们尝试使用基于半监督学习的推理方法，即同时考虑少量的标记数据和大量的无标记数据，从而解决了前述的标记数据少且获取困难的问题，具有重要意义。半监督学习主要有基于约束条件和基于距离度量的半监督学习，同时还有基于模型和基于密度的半监督学习，以及基于数据集空间结构方法的半监督学习方法[17~19]。Catanzaro 等人[17]提出了将半监督谱学习与隐马尔科夫模型相结合的人脸识别算法，解决了标记信息相对较少的问题。Mireille[18]讨论了基于密度的半监督学习方法，从约束关系着手，拓展 must-link 和 cannot-link 关系集合，以满足即使只有部分标记信息情况下，依然可以很好地训练学习。

但是以上半监督学习方法针对类不均衡的训练数据均未作深入讨论。类不均衡数据并非传统意义上的噪声数据，而是广泛存在于异常检测、医疗诊断等各个领域中的真实存在的数据集合，这些数据中的一类属于正常数据，容易获得其标记信息，而另一部分则由于数据存在相对概率小、数据特征难以捕捉等各种原因，导致了类不均衡情况的出现。类不均衡数据主要有数据稀缺性、将数据分类预测倾向于多类、不平衡数据难以识别以及决策面偏移失衡等问题。现在研究的很多聚类分类问题大都是在类别大致平衡的条件下讨论的，因而对于类不均衡数据往往得不到有效的处理。类不均衡问题容易导致划分面的位置过度偏向于一类，可能最终结果是基本上把几乎所有的正类和负类都划在划分面的一侧，使得最终的结果都为同一类，甚至将正常数据错划分为噪音数据。

针对以上问题，本文提出了类不均衡的半监督高斯过程分类算法，其基本原理是利用数据空间分布的自适应特性，利用极少量不平衡的标记数据来构建半监督分类器，用其对未标记数据进行分类，从而使数据的分类标签信息达到相对平衡。该算法中，未标记数据通过半监督自训练的方法逐渐被标注上类别标记，并且被加入到训练样本集合中作为标记数据，从而可进行新的训练。半监督分类器不断通过自我训练，获取可信度高的标记数据来平衡数据中原来存在的类别不均衡问题。对未标记数据进行类别标记是迭代进行的，通过判断设置的阈值或者迭代次数，决定是否进行重复训练直至达到要求为止。半监督高斯过程分类器可以根据要求主动寻找数据内部的类别信息进行自动分类，训练自动化，减少了人工标记的错误率，提高了分类标记数据平衡比例和数量，从而解决了类不均衡数据稀缺而导致的错分问题，提高了对不平衡数据进行正确分类的准确度，算法简单而高效。基于自适应类不均衡的半监督高斯过程分类算法的具体步骤是：首先根据类不均衡数据特性进行数据预处理，然后利用少量的标记数据进行高斯过程分类训练，选取预测概率置信度最高的未标记数据，向该未标记数据注入合理的类标记信息，并且自动地将新标记过的数据样本加入到原有的训练集中，用扩充后的训练集再次进行高斯过程分类；最后采用自训练迭代执行，直至构造出最优的高斯过程分类器，用以对测试数据集分类。算法将标记数据与未标记数据结合，实现自训练的半监督高斯过程分类，即使在有少量标记数据训练样本的情况下，同样保证了分类结果的准确度。本文通过多组实验对类不均衡的半监督高斯过程分类算法的效果进行了验证。

2　基础工作

高斯过程分类算法的核心思想是：把非高斯的真实后验分布 $p(f|D,\theta)$通过一个高斯类近似后验分布 $q(f|D,\theta)=N(f|(\mu,\sum))$来代替，再通过此后验分布给出测试数据的近似预测分布。其中，μ 为均值，$\sum$表示方差。

对于高斯过程分类(GPC)问题的定义：给定的训练数据集 $D=\{(x_i,y_i), i=1,2,\cdots,m\}$，$x_i$ 为连续数据，表示特征向量，y_i 为离散数据，表示类别标记。目标是对于新输入 x^*，预测其输出 y^*。若 y 取值为{0,1}或者{1,−1}称为二类分类；若 y 取值为多个整数值，称为多分类。本文主要讨论二分类问题。

对于确定的输入矢量 $\boldsymbol{x}$，$p(y|\boldsymbol{x})$分布为伯努利分布，y=1 的概率为 $p(y=1|\boldsymbol{x})=\Phi(f(\boldsymbol{x}))$，其中，$f(\boldsymbol{x})$称为潜在函数，服从高斯过程：$f(\boldsymbol{x}|\theta)\sim GP(0,K)$。$f(\boldsymbol{x})$定义了标记数据集合和相对应的类标记集 Y 之间的映射关系。Φ 函数为标准高斯分布的累积概率密度函

数，取 Sigmoid 函数，保证概率值落在[0,1]区间。

由于给定的潜在函数 f，其观测数据是相互独立的伯努利分布变量，似然函数可以描述为

$$p(y \mid f)=\prod_{i=1}^{m} p(y_i \mid f_i)=\prod_{i=1}^{m} \varphi(y_i \mid f_i) \tag{1}$$

潜在函数的先验分布为

$$p(f \mid X,\theta)=N(0,K) \tag{2}$$

在式(2)中，K 定义了协方差矩阵(核函数)，$K_{ij}=k(x_i,x_j,\theta)$，这里 $k(\cdot)$是与 θ 有关的正定协方差函数；θ 可通过潜在函数 f 的极大似然法来估计[20]得到最优超参数。

高斯过程模型的协方差函数需要满足：对任一点集都能够保证产生一个非负正定协方差矩阵。本文采用的协方差函数为高斯核函数。

$$K(\| x-x_c \|)=\sigma_f^2 \exp\left\{-\frac{1}{2l^2}(x-x_c)^2\right\} \tag{3}$$

其中，x_c 为核函数的中心，超参数 $\theta=\{\sigma_f, l\}$。由式(3)可以看出，协方差函数由 2 部分组成：第一部分用来表示 2 个数据点间的距离相关性，如果它们的距离相对于宽度参数 l 很小，即相关性高，指数项就趋于 1；否则两数据点之间相关性低，指数项就趋于 0。超参数 σ_f用来控制局部相关性的程度。

当获得实际观察值后，根据贝叶斯规则，潜在函数 f 的后验分布为

$$\begin{aligned} p(f \mid D,\theta) &= \frac{p(y \mid f)p(f \mid X,\theta)}{p(D \mid \theta)} \\ &= \frac{N(0,K)}{p(D \mid \theta)}\prod_{i=1}^{m}\Phi(y_i f_i) \end{aligned} \tag{4}$$

GPC 模型的主要目的是对于给定的测试输入 x^*，预测其所属的类别 y^*。给定测试数据点 x^*后，与 x^*对应的潜在函数值 f^*的条件概率为

$$p(f^* \mid D,\theta,x^*)=\int p\left(y^* \mid f,X,\theta,x^*\right)p\left(f \mid D,\theta\right)\mathrm{d}f \tag{5}$$

故 x^*的类标记预测概率 y^*为

$$p(y^* \mid D,\theta,x^*)=\int p(y^* \mid f^*)p(f^* \mid D,\theta,x^*)\mathrm{d}f \tag{6}$$

将近似高斯后验分布代入式(5)中，可得到潜在函数 f^*在测试数据点 x^*的近似高斯后验分布。

$$q(f^* \mid D,\theta,x^*)=N(f^* \mid (\mu^*,\sigma_*^2)) \tag{7}$$

其中，均值和方差为

$$\mu^*=k^{*\mathrm{T}}K^{-1}\mu \tag{8}$$

$$\sigma^{*2}=k\left(x^*,x^*\right)-k^{*\mathrm{T}}\left(K^{-1}-K^{-1}AK^{-1}\right)k^* \tag{9}$$

其中，$k^*=[k(x_1, x^*),\cdots,k(x_m, x^*)]^{\mathrm{T}}$ 表示测试数据 x^*与训练数据集的先验协方差函数。

3 类不均衡半监督高斯过程分类算法

3.1 相关定义

本文将半监督学习思想与高斯过程机器学习相结合，综合利用类不均衡数据特点进行半监督训练，提出了类不均衡的半监督高斯过程分类算法。对于给定的训练数据集，将其中的一小部分数据定义为标记数据对象，其他数据为未标记数据对象。下面给出具体的相关定义。

定义 1 令 X 表示数据对象集合，X^{L} 表示该集合中的原始标记数据集，X^{U} 表示未标记数据集，则 $X=\{X_1X_2\cdots X_n\}$, $X^{\mathrm{L}}=\{X_1^{\mathrm{L}}\ X_2^{\mathrm{L}}\cdots X_p^{\mathrm{L}}\}$, $X^{\mathrm{U}}=\{X_1^{\mathrm{U}}\ X_2^{\mathrm{U}}\cdots X_q^{\mathrm{U}}\}$，其中，$n$ 表示数据集数目，p 表示标记数据集数目($1\leqslant p<n$), q 表示为未标记数据集数目，并且满足 $n=p+q$，即 $X=X^{\mathrm{L}}+X^{\mathrm{U}}$。

定义 2 令 Y 表示标记数据集合的类标记信息，$Y=\{y_1, y_2, \cdots, y_p\}$，$y_i\in\{1, -1\}$，$Y$ 与 X^{L} 数据集中的元素一一对应。

定义 3 预测概率置信度 α，在进行半监督高斯过程分类训练时，若预测概率达到置信度 α 或者小于 α 则将该数据考虑是否注入类标记。α 可人工设置，根据多次实验经验，本文设定 α=0.95 为最佳置信度阈值。

3.2 算法描述

类不均衡的半监督高斯过程分类算法主要结合高斯过程分类算法和自训练半监督学习方法以解决类不均衡数据分类问题。该算法包括引用文献[21]的部分算法 1 和本文提出的算法 2 两部分，具体描述如下。

算法 1 高斯过程分类(GPC)算法[21]

输入：协方差矩阵 $\boldsymbol{K}$，训练集标记 Y，似然函数 $p(y \mid f)$；

输出：分类预测分类函数 f。

Step1 初始化预测函数 f=0。

Step2 令对角矩阵 $\boldsymbol{W}=-\nabla\nabla\log p(y \mid f)$，对矩阵 $\boldsymbol{L}$ 做 cholesky 矩阵分解，使得

$$\boldsymbol{L}=cholesky(I+\boldsymbol{W}^{1/2}K\boldsymbol{W}^{1/2})$$

Step3　计算 $b=\boldsymbol{W}f+\nabla\log p(y\mid f)$，

$a=b-\boldsymbol{W}^{1/2}L^{\mathrm{T}}\backslash(L\backslash(\boldsymbol{W}^{1/2}Kb))$。

Step4　$f=Ka$，若超过迭代次数或目标函数收敛转 Step5，否则转 Step2。

Step5　计算似然函数 $\log q(y\mid X,\theta)=-\frac{1}{2}a^{\mathrm{T}}\cdot f+\log p(y\mid f)-\sum_i\log L_{ii}$。

Step6　返回 f 和 $\log q(y\mid X,\theta)$，算法结束。

算法 1 为高斯过程二分类构造器构造过程[21]。通过目标函数建立收敛准则。f 是由牛顿计算公式而得到的隐变量的最大后验概率，即分类预测函数，$\log q(y\mid X,\theta)$ 是边缘最大似然函数，可以通过 f 和矩阵 $\boldsymbol{W}$ 不断地对其进行优化，使该函数通过训练数据样本低密度区域最终得出分类预测函数。

算法 2　半监督高斯过程分类(SSGP)算法

输入：标记数据集合 X^{L}，未标记数据集合 X^{U}，测试数据集 X^{T}。

输出：分类预测结果 R，R 与测试集 X^{T} 中一一对应，且 $r_i\in\{1,-1\}$。

Step1　标记数据集 X^{L} 全部复制到 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 中。

Step2　将更新后的标记数据集合 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 作为训练集，利用算法 1 输入到高斯过程，进行分类训练学习，构造高斯过程分类器。

Step3　使用 Step 2 构造的高斯过程分类器，对未标记数据集合进行自训练分类，且对分类结果做如下筛选：若选取的数据点 X_i 预测概率 $p\geqslant\alpha$，则将该数据点 X_i 加入到 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 中，置其类标记信息 Y_i=+1；若选取预测概率 $p\leqslant1-\alpha$，则将该数据点 X_i 加入到 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 中，置其类标记信息 Y_i=−1；同时从 X^{U} 去除该数据信息。更新标记数据集 $X^{\mathrm{L}}{}_{\mathrm{new}}$，未标记数据集合 X^{U}；

Step4　若更新后的 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 与 X^{L} 数据集相同或者构造的分类器分类结果稳定即概率相同时，停止 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 的更新，输出数据集 $X^{\mathrm{L}}{}_{\mathrm{new}}$ 和对应的标记信息 Y；否则，置 X^{L}=$X^{\mathrm{L}}{}_{\mathrm{new}}$，更新 X^{L}，重复 Step2。

Step5　更新后的 X^{L} 为训练集合，利用算法 1 构造高斯过程分类器，对分类数据集合 X^{T} 进行分类，输出 X^{T} 的类标记信息 R，算法 2 结束。

算法 2 通过自训练的半监督学习方法向未标记数据注入类标记信息，用扩充后的标记数据集构造分类器，未标记数据反馈预测结果指导下次的分类。

3.3　算法复杂度

在 SSGP 算法的自训练过程中，利用少量的标记数据进行高斯过程分类训练，得到一个初始学习器，然后选取预测概率置信度最高的未标记数据注入标记，同时将新标记的样本加入到原来的训练集中，随后使用这个扩充后的训练集再次进行高斯过程分类，重新训练学习器，重复以上过程直到满足迭代终止条件。算法从无标记数据和有标记数据开始，通过将无标记样本整合进有标记样本中，自训练在这个过程中实际上进行的是一个强化过程，目的是为了改进学习器性能。

SSGP 算法的复杂度与 GPC 算法的复杂度紧密相关，但是由于 GPC 算法用不同的方法进行优化近似求解，其时间复杂度和空间复杂度差异较大，因此不容易直接计算 SSGP 算法的复杂度[22]。根据文献[23]可以通过计算 SSGP 算法执行时所用的训练样本总数来衡量算法的复杂度。定理 1 表明，SSGP 算法与标记样本和未标记样本的数量呈线性关系而不是指数关系。

定理 1　SSGP 算法执行时所用的训练样本复杂度为 $O(Max\_Iter(p+q))$，其中，Max_Iter 是半监督训练最大迭代次数，p 和 q 分别是标记样本集数目和未标记样本集数目。

证明　设 s 为从未标记样本集 X^{U} 中选择出来的置信度最高的样本所占的比例，在第一次迭代训练中，p 和 q 分别是标记样本集数目和未标记样本集数目。在第一次迭代后，qs 个未标记样本被标记，并加入到已标记样本中，且有 qs 个样本从 X^{U} 中删除。因此，在第二次迭代训练中，已标记样本数为 $p+qs$，未标记样本数为 $q(1-s)$。依此类推，在第 i 次迭代训练中，已标记样本数为

$$\begin{aligned}&p+qs+qs(1-s)+\cdots+qs(1-s)^{i-2}\\&=p+q(1-(1-s)^{i-1})\end{aligned}\tag{10}$$

未标记样本数为 $q(1-s)^{i-1}$。由此可以得出，训练样本总数为

$$\begin{aligned}&\sum_{i=1}^{Max\_iter}(p+q(1-(1-s)^{i-1}))\\&=Max\_iter(p+q)-q\cdot\frac{1-(1-s)^{Max\_iter}}{s}\end{aligned}\tag{11}$$

因为

$$Max\_iter(p+q)-q\cdot\frac{1-(1-s)^{Max\_iter}}{s}<Max\_iter(p+q)$$

所以 SSGP 算法执行时所用的训练样本复杂度为 $O(Max\_Iter(p+q))$，定理得证。

4 实验结果与分析

4.1 仿真数据集实验分析

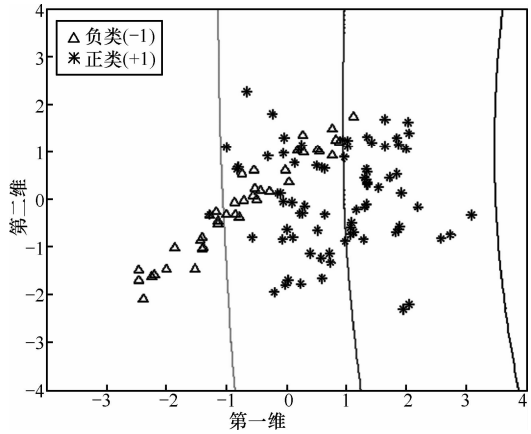
为验证 SSGP 算法对数据集信息具有更好的提取，本文首先将仿真数据进行实验对比。仿真数据是从 2 个不同的二维正态分布中随机采样 80、40 个数据点，共 120 个数据点。图 1 为用 GP 算法与 SSGP 算法时，仿真数据信息的边缘似然值和空间分布信息。

图 1 说明了高斯过程使用 SE 内核在二分类问题上的分类情况，SE 内核函数为一个 variable length-scale 和 logistic 响应函数。Laplace 近似法画出该数据集的似然函数值，体现该数据信息。这些数据点分布在[−4,4]，似然值则集中在[0,1]区间范围内。*和 Δ 分别表示 2 类数据，从图 1 中可以很明显地区分出每一类。图 1 显示了二维数据点分布空间情况。这些数据点被分成 2 类，*代表正类(+1)，Δ 代表负类(−1)，图 1 中等高线为不同超参数情况下的预测概率等高线，越是接近于 1 的等高线，其值极有可能被分为正类，相反，接近于 0 的等高线一般被划分为负类。

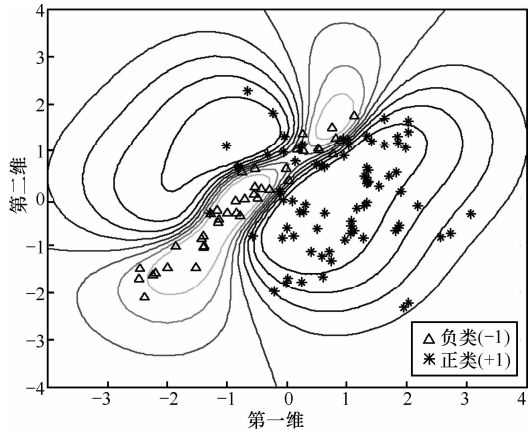
图 1(a)表示为未优化情况，数据错分情况十分严重，难以正确分类，显示的几条等高线都是在 0.2~0.5 之间，基本上无法分类。图 1(b)显示通过 GP 算法优化后，其新的超参数对构造分类器更加合理些，但是由于部分等高线仍然是从高密度区域穿过，依然存在错分问题。在图 1(c)中，SSGP 算法在正确划分数据的基础之上，提高了可信度，似然函数等高线尽可能地从低密度区域划分，数据集中在 0.1 或 0.9 附近。实验表明，经过超参数优化后的高斯过程算法尽可能地平衡了 length-scale，使似然函数等高线从低密度区域穿过，提高了分类准确度，减少了错分数，增加了可信度。

4.2 USPS 数据集实验

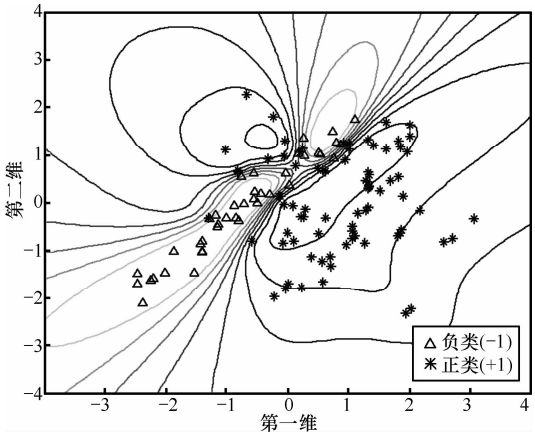
为验证类不均衡的半监督高斯过程分类算法的有效性和可行性，本文采用 USPS 手写数据集进行性能测试对比实验。在 USPS 数据集中，共有 9 298 个 16×16 灰度图像，经过数据预处理，其像素强度均在[−1,1]范围内。从该数据集中，笔者提取了数字“3”和“5”，其中，训练样本 767 个，包括 406 个“3”和 361 个“5”。测试样本 773 个，包括 418 个“3”和 355 个“5”。在本组实验的训练样本中，只选取其中一部分样本用于训练，并将正类标记与负类标记比例依次从 1:1~1:16 做多组比较。



(a) 超参数 σ=3.0，l=0.0



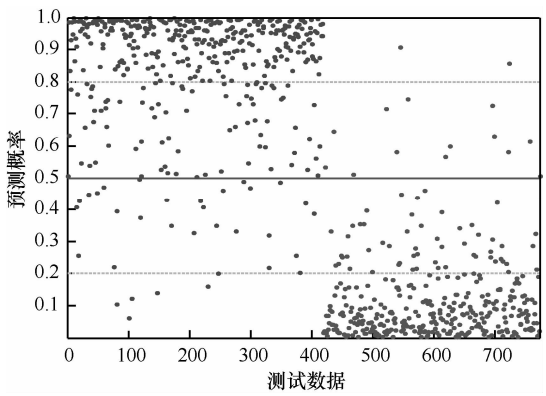
(b) 超参数 σ= 0.183 3，l=2.511 6



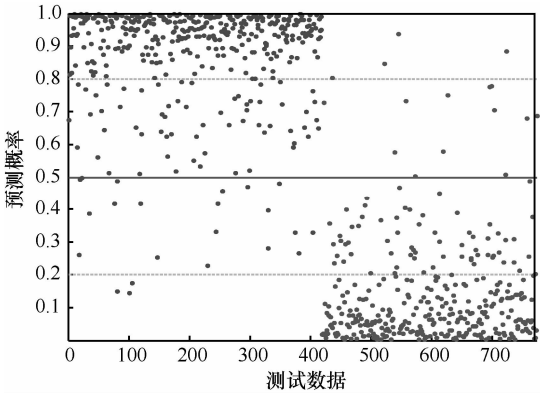
(c) 超参数 σ=−1.039 9，l = 1.975 7

图 1 不同超参数下的似然函数

图 2 和图 3 实验结果分别为标记比率为 1:1 和 1:16 条件下，其 GP 算法与 SSGP 算法性能比较。图 2 和图 3 所示为分类预测概率示意，从中可以看

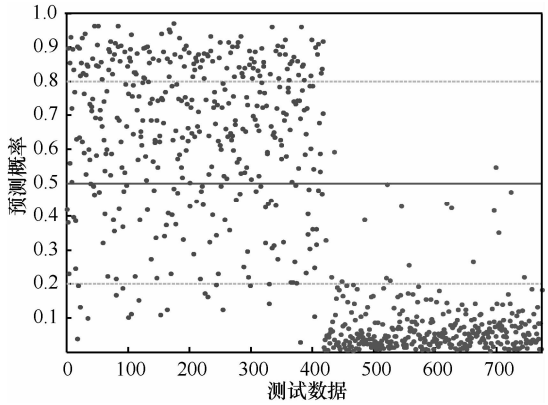


(a) GP 算法的预测概率

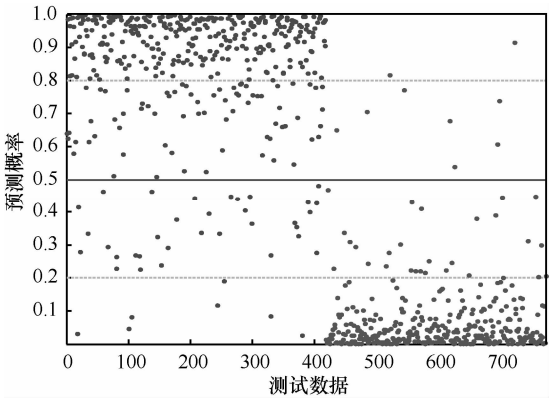


(b) SSGP 算法的预测概率

图 2　类标记比率 1:1 的预测概率散点对比



(a) GP 算法的预测概率



(b) SSGP 算法的预测概率

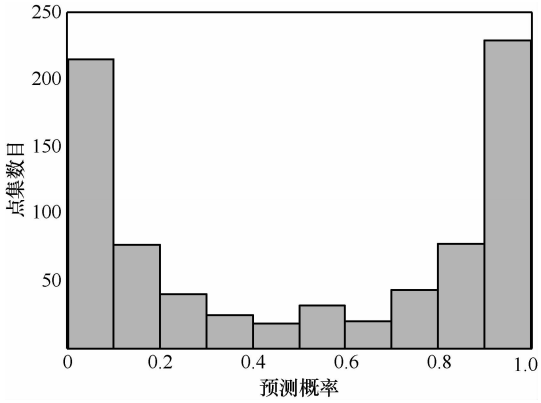
图 3　类标记比率 1:16 的预测概率散点对比

出半监督学习高斯过程算法的预测概率明显优于高斯过程算法。在 1:1 条件下，正负类比率相同，GP 预测精度为 93.79%，SSGP 算法利用半监督学习得到的信息将预测精度提高到了 95.08%，而正类的预测准确度分别为 92.11%和 94.74%，也都达到了很高的准确度。实验表明，在高斯过程分类算法中，当训练集标记比率接近时，单类预测概率接近于整体预测概率，且标记信息较多的情况下，预测准确度很高，SSGP 算法的性能与 GP 算法性能效果相似，半监督信息利用率较低。在 1:16 条件下，正类标记数大大降低，训练集中大部分都是负类标记信息。此时，GP 算法对全局数据预测精度为 86.55%，而 SSGP 算法对全局数据预测精度为 93.66%，提高了 7.11%，SSGP 算法的优越性逐渐体现出来了。对于正类的预测分类，GP 算法和 SSGP 算法准确度分别为 75.12%和 88.52%，虽然都不如比率为 1:1 情况预测分类的准确度高，但 GP 算法降低了 16.99%，而 SSGP 算法仅仅降低了 6.22%。此外，在数据失衡为 1:16 的情况下，SSGP 算法比 GP 算法在正类准确度上提高了 13.4%。由此可见，SSGP 算法在数据比率严重失衡的情况下，充分利用半监督信息，扩充了标记信息集合，提高了分类准确度，其算法依然具有较好的稳定性、顽健性，性能幅度下降尽可能地小。而高斯过程分类算法的分类准确性虽然很好，但是难以适应真实数据情况，一般都是在理想数据集下的性能比较，未充分考虑到数据的各种情况，在数据失衡情况下，其预测分类性能低于 SSGP 算法的分类性能。实验表明，在高斯过程分类算法中，当训练集标记比率接近时，且标记信息较多的情况下，预测准确度很高，SSGP 算法的性能与 GP 算法性能效果相似，半监督信息利用率较低。当训练集标记比例严重失衡时，全局预测依赖于单类预测结果，训练数目较少的一类极有可能被训练数目较大的一类所覆盖，造成该类预测概率严重降低，错分数目大于各种情况。此时 GP 算法不再完全适用该情况，而 SSGP 算法则可以利用数据集中的未标记信息指导高斯过程分类，通过可信度判断不断地进行标记信息的

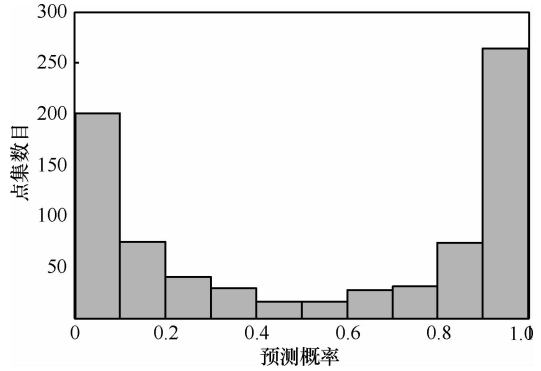
扩展，从而提高分类器的准确度，为分类预测数据提供更加可靠的分类器和精确度。

实验表明，在数据失衡条件下，高斯过程算法的预测概率除了错分较多之外，密度分布还不明显，很多预测概率点都落在了 p=0.5 附近，难以辨析其分类结果，预测概率低，准确度不高。类不均衡的半监督高斯过程分类算法通过对未标记数据的学习，训练集获取的已知信息量增大，构造分类器精度提高，概率密度分布集中且主要集中在概率 1 和概率 0 附近，即对数据点分类更加明确，不确定性减小，准确度大大提高。分析实验可知，使用自训练半监督学习的高斯过程算法在进行分类器训练时更加有效，进行分类预测时更加准确。SSGP 算法分类的确定性和预测的稳定性明显优于高斯过程分类算法。

图 4 和图 5 分别表示在不同标记比率下，使用 GP 算法与 SSGP 算法时的分类可信度情况。实验规定，预测概率越接近于 1 和 0 时，预测准确度越高，可信度越大。比率为 1:1 时，GP 算法和 SSGP 算法预测概率值都主要集中在 0 和 1 附近，表明分类的可信度很高，准确度也相应较高，且两者区分不大，

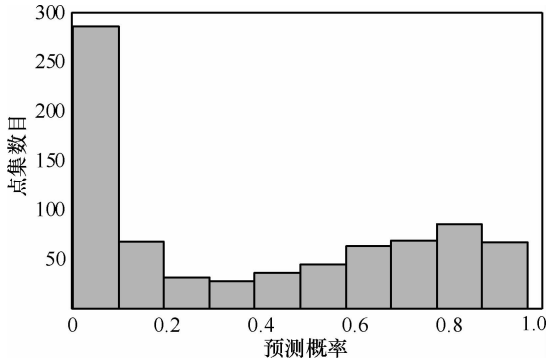


(a) GP 算法的数据统计

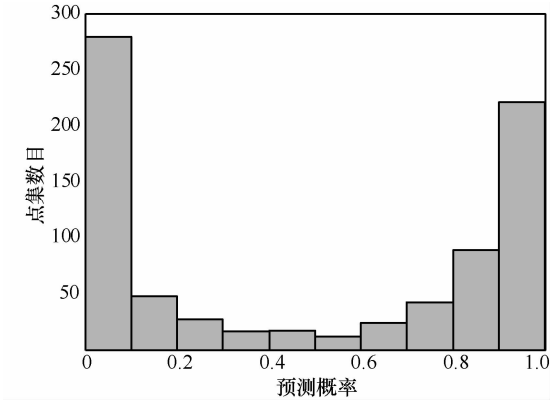


(b) SSGP 算法的数据统计

图 4 类标记比率 1:1 的概率统计柱状示意



(a) GP 算法的数据统计



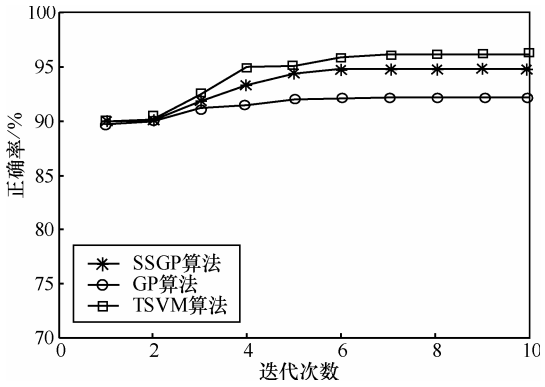
(b) SSGP 算法的数据统计

图 5 标记比率 1:16 的概率统计柱状示意

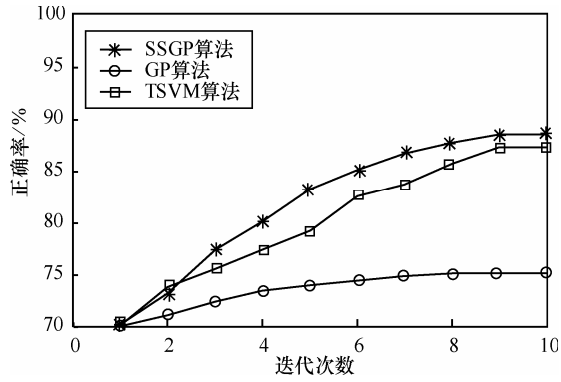
SSGP 算法此时可以认为近似于 GP 算法。而比率为 1:16 时，对于标记信息多的负类，影响不大，仍然集中在 0 附近，可信度依然很高，而对于笔者想得到的正类分类情况，利用 GP 算法却难以获得相应的正确分类，具有较低的可信度，概率 1 附近几乎没有数据，SSGP 算法却可以根据未标记信息，进行迭代计算后，可信度逐渐提高。图 5(b)显示了与图 4 近似的效果。在该组实验中，标记比例依次从 1:1~1:16，标记数据逐渐失衡，GP 算法分类预测效果逐渐不佳，当最后为 1:16 时，其正类可信度大大下降。而 SSGP 算法克服了数据失衡的问题，保持了分类的可信度和准确度。实验表明，SSGP 算法具有更加稳定的可靠性和有效性。

图 6 为不同标记比率下，GP 算法、TSVM 算法与 SSGP 算法在同一数据集上的迭代次数与准确度的关系。在数据比例为 1:1 的情况下，3 种算法的分类准确度曲线都相对较为平缓，预测分类也都比较高。从图 6(a)可以发现，TSVM 分类算法的分类精度较高，分类效果较好，SSGP 算法分类准确度略低于 TSVM 算法，但 3 种算法分类准确度都达到了 90%以上，无太大差异。而在 1:16 的情

况下，SSGP 算法在进行第 4 次迭代后，准确度明显攀升较快，分类性能较佳，效果相对明显。TSVM 算法虽然分类效果也较好，但相对于 SSGP 分类算法准确度低了 1.2 个百分点，在迭代过程中，其准确度也一直低于 SSGP 算法的分类效果。从图 6(b)可以看出，随着比例逐渐失衡，SSGP 算法准确度一直保持相对较高水平，更加适合数据比例失衡的分类。



(a) 标记比例 1:1



(b) 标记比例 1:16

图 6　GP 算法与 SSGP 算法在不同比例下的性能曲线（正类准确度）

表 1 主要讨论了数据在不同标记比例失衡情况下，SSGP 算法与 GP 算法、TSVM 算法分类精度的问题。经过对比可发现，若标记比例为 1:1 时，两者精度差异不明显，但随着标记比例的逐步增加，传统的 GP 算法难以应付，其中一类分类错误率大大提高，其可信度也在不断地降低，TSVM 算法准确度也由原来的 96.06%下降到了 87.32%，波动幅度较大。而 SSGP 算法通过自训练得到部分标记数据，提高了标记数目，虽然调整了标记比例，其最终效果还是得到了相应的提高，尤其对标记比例相对比较少的一类，效果尤为明显，可信度也提高了很多，对整体分类的准确度有了较大的改善。从实验中可以分析得出，在标记比例严重失衡的情况下，SSGP 算法有更高的准确度和可信度。实验显示，为了达到相同的预测分类准确率，经典高斯过程算法需要使用更多的标记数据，这在实际需求中将会增加相应代价，也说明了自训练的半监督高斯过程分类算法在数据失衡的情况下确实能起到提高预测准确率的作用。由此得到下述结论：在具有少量标记数据信息或标记信息不对称情况下，较 GP 算法与 TSVM 分类算法而言，SSGP 算法能更充分利用少量标记数据进行数据分类。

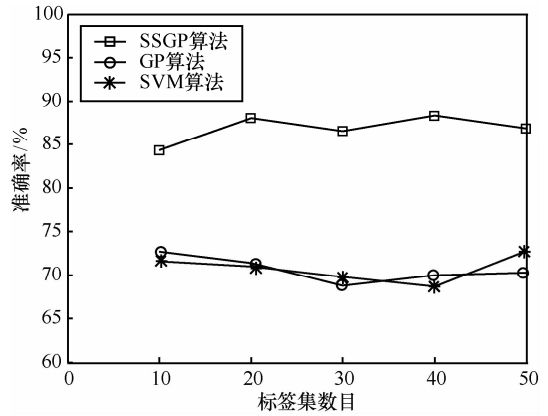
表 1　GP 算法、TSVM 算法与 SSGP 算法在不同比例的性能比较

标记比例	算法的全局准确度/%			算法的正类准确度/%		
	GP	TSVM	SSGP	GP	TSVM	SSGP
1:1	93.79	96.21	95.08	92.11	96.06	94.74
1:2	93.53	94.35	94.75	90.43	92.71	93.54
1:4	91.37	93.98	94.69	86.12	89.56	91.63
1:8	90.03	93.57	94.82	82.30	88.74	89.95
1:16	86.55	93.24	93.66	75.12	87.32	88.52

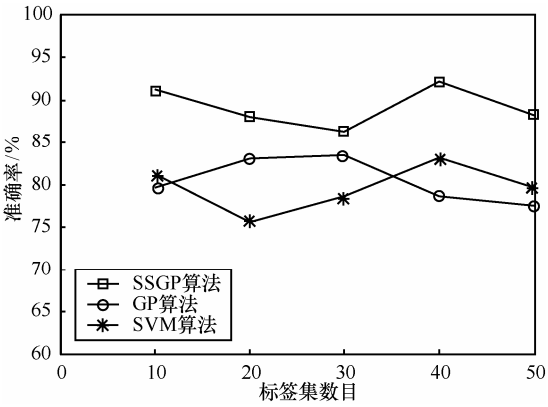
4.3　极端数据比例失衡实验

为进一步研究标记数据率对分类结果的影响并验证本文提出的 SSGP 算法的有效性，分别从训练样本数据类别比例失衡和数据标记率失衡角度进行实验研究。本文在 USPS 数据集上抽取了“2”、“3”和“5”共 3 种数据构成实验数据集，并且比较了 3 种标记数据率：1/30、1/20、1/10。假设数字“3”为正常数据类(正类)，数字“2”和“5”的混合集为异常数据类(负类)，正负类数据量比例为 20:1，每次实验都在该标记数据率的情况下，随机选取标记数据集 50 次，总共进行 150 次对比实验。比较了 GP 算法与 SSGP 算法及 SVM 算法的平均性能、最佳性能和最差性能，实验结果如图 7 和表 2 所示。

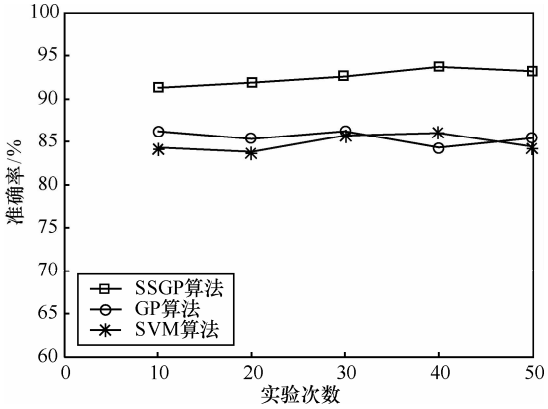
图 7 为 SSGP 算法、GP 算法和 SVM 算法的数据分类情况，在训练样本情况较少、比例严重失衡的条件下，自训练的半监督高斯过程分类算法充分利用未标记信息，通过学习获得了更多的标记信息，增加了算法的准确度。尤其对于负类，该算法通过从测试集的未标记信息中扩展负类标记信息，尽量使得负类标记信息数据分布与整体负类信息数据分布相似，为构造分类器提供了更多的信息参考，



(a) 标记率为 1/30



(b) 标记率为 1/20



(c) 标记率为 1/10

图 7　SSGP 算法与 GP 算法、SVM 算法在不同标记数据率下性能对比

明显提高了每一类的数据分类准确度。实验结果显示，在 1/30 的标记数据率情况下，SSGP 算法平均错分率为 12.57%，而 GP 算法和 SVM 算法平均错分率分别为 29.40%和 29.57%。经过多次随机实验表明，即使在最佳情况下，GP 算法和 SVM 算法依然难以达到 SSGP 算法的平均水平。随着标记数量的增加 SSGP 算法、GP 算法和 SVM 算法的性能均得到了相应的提升，从图 7 和表 2 中可以看出，在 1/20 和 1/10 两种标记比率下，GP 算法与 SVM 算法性能提高幅度较大，也就是说标记数据率的提高使得这 2 种算法的分类期望和均差均得到了很大的提高。由此可以得到下述结论：在具有少量标记数据信息下，较高斯过程分类和 SVM 算法而言，自训练的半监督高斯过程分类算法更能充分利用少量标记数据来指导未标记数据进行分类。此外，从图 7 中还可以看出，选取标记数据存在很大的随机性和误差，如果开始选取的标记数据处在分布边缘，则分类效果不明显。然而使用半监督高斯过程分类算法时，首先在分类之前进行二次选择，构造更新的标记数据集，约束了标记数据的选取。其次，通过半监督学习方法向未标记数据中注入类标记，使得最后的标记数据分布与全局数据分布的相似度提高，减小了边缘值对整体数据分类的影响，有效地克服了初始随机选择标记数据带来的不良影响，提高了分类精度。

表 2　3 种算法在 USPS 数据集不同标记数据上的性能对比

标记数据率	SSGP 错分率/%			GP 错分率/%			SVM 错分率/%		
	平均	最佳	最坏	平均	最佳	最坏	平均	最佳	最坏
1/10	7.48	6.33	8.72	14.51	13.74	15.66	15.18	13.96	16.26
1/20	10.8	7.86	13.64	19.30	16.48	22.46	20.33	16.79	24.36
1/30	12.57	11.63	15.74	29.40	27.32	31.29	29.57	27.57	31.63

5　结束语

本文提出了一种类不均衡的半监督高斯过程分类算法，利用未标记数据集合进行半监督学习，把未标记数据集通过半监督学习将其部分转换为标记数据，有效地解决了高斯过程分类训练中标记数据集过少、类不均衡的问题，增加了高斯过程在异常数据分类中的准确性，提高了其分类精度和可信度。实验结果表明了该算法的可靠性和有效性。

参考文献：

[1] KITAYAMA S, YAMAZAKI K. Simple estimate of the width in Gaussian kernel with adaptive scaling technique[J]. Applied Soft Computering, 2011, 11(8):4726-4737.

[2] RODNER E, WACKER E S, KEMMLER M, *et al*. One-class classification for anomaly detection in wire ropes with Gaussian processes in a few lines of code[A]. Proceedings of the 12th IAPR Conference on Machine Vision Applications (MVA)[C]. Nara, Japan, 2010. 296-308.

[3] 姚伏天.基于高斯过程的高光谱图像分类研究[D].杭州:浙江大学, 2011.
YAO F T. Gaussian Processes based Classification for Hyperspectral

Imagery[D]. Hang Zhou: Zhejiang University, 2011.

[4] KAPOOR A, GRAUMAN K, URTASUN R, *et al*. Gaussian processes for object categorization[J]. International Journal of Computer Vision, 2010, 88(2):169-188.

[5] 孙欣尧,王雪,王晟.无线传感网络协同概率多模识别方法[J]. 通信学报, 2011, 32(6):141-147.
SUN X Y, WANG X, WANG C. Collaborative probability based multimodel target identification in wireless sensor networks[J]. Journal on Communications, 2011, 32(6):141-147.

[6] 熊志化. 高斯过程模型及其在工业过程软测量中的应用研究[D]. 上海: 上海交通大学,2006.
XIONG Z H. Study on Gaussian Process Model and Its Application to Soft Sensor in Process Industries[D].Shanghai: Shanghai Jiao Tong University, 2006.

[7] VAN GOOL E, WINN W, ZISSERMAN A. The PASCAL visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2):303-338.

[8] 陈凤. 基于 HRRP 和 JEM 信号的雷达目标识别技术研究[D].西安: 西安电子科技大学,2009.
CHEN F. Radar Target Recognition Based on HRRP and JEM Signal[D]. Xi'an: XiDian University,2009.

[9] 王磊, 邹北骥, 彭小宁等. 基于高斯过程的表情动作单元跟踪技术[J]. 电子学报, 2007, 35(11):2087-2091.
WANG L, ZOU B J, PENG X N, *et al*. Facial tracking by Gaussian process[J]. Acta Electronica Sinica, 2007, 35(11):2087-2091.

[10] DEISENROTH M P, TURNER R D, HUBER M F, *et al*. Robust filtering and smoothing with Gaussian processes[J]. IEEE Transactions on Automatic Control, 2012, 57(7):1865-1871.

[11] GASBARRA D, SOTTINEN T, ZANTEN H V. Conditional full support of Gaussian processes with stationary increments[J]. Journal of Applied Probability, 2011, 48(2):561-568.

[12] RODNER E, DENZLER J. One-shot learning of object categories using dependent Gaussian processes[A]. Proceedings of the DAGM Conference on Pattern Recognition[C]. Springer, Heidelberg, 2010. 232-241.

[13] BOSCH A, ZISSERMAN A, MUNOZ X. Representing shape with a spatial pyramid kernel[A]. ACM International Conference on Image and Video Retrieval (CIVR)[C]. Amsterdam, Netherlands, 2007. 401-408.

[14] CHUM O, ZISSERMAN A. An exemplar model for learning object classes[A]. ACM International Conference on Image and Video Retrieval (CIVR)[C]. Amsterdam, Netherlands, 2007. 19-21.

[15] HAGERW W. Updating the inverse of a matrix[J]. Society for Industrial and Applied Mathematics (SIAM) Review, 1989, 31(2):221-239 .

[16] ADANKON M M, CHERIET M. Model selection for the LS-SVM application to handwriting recognition[J]. Pattern Recognition, 2009, 42(12):3264-3270.

[17] CATANZARO B, SUNDARAM N, KEUTZER K. Fast support vector machine training and classification on graphics processors[A]. Proceedings of the 25th International Conference on Machine Learning(ICML)[C]. New York, NY, USA, 2008. 104-111.

[18] TOHME M, LENGELLE R. Maximum margin one class support vector machines for multiclass problems[J]. Pattern Recognition Letters, 2011,32(13):1652-1658.

[19] FENG W, XIE L, ZENG J, *et al*. Audio-visual human recognition using semi-supervised spectral learning and hidden Markov models[J]. Journal of Visual Languages & Computing, 2009, 20(3):188-195.

[20] RUIZ C, SPILIOPOULOU M, MENASALVAS E. Density-based semi-supervised clustering[J]. Data Mining and Knowledge Discovery, 2010, 21(3):345-370.

[21] RASMUSSEN C E, WILLIAMS C K I. Gaussian Processes for Machine Learning[M]. Cambridge: MIT Press, 2006.

[22] 陈晓峰, 王士同, 曹苏群. 半监督多标记学习的基因功能分析[J]. 智能系统学报, 2008, 3(1):83-90.
CHEN X F, WANG S T, CAO S Q. Gene function analysis of semi 2 supervised multi-label learning[J]. CAAI Transactions on Intelligent Systems, 2008, 3(1):83-90.

[23] KLAUS B, JOHANNS F, EYKE H. A unified model for multilabel classification and ranking[A]. Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence[C]. Riva del Garda, Italy, 2006. 489-493.

作者简介：

夏战国（1974-），男，河北保定人，中国矿业大学博士生，主要研究方向为模式识别与人工智能。

夏士雄（1961-），男，辽宁抚顺人，中国矿业大学教授、博士生导师，主要研究方向为矿山智能信息处理。

蔡世玉（1988-），男，安徽定远人，中国矿业大学硕士生，主要研究方向为机器学习与数据挖掘。

万玲（1988-），女，宁夏石嘴山人，中国矿业大学硕士生，主要研究方向为智能信息处理。