文章编号:1001 - 9081(2009)06 - 1578 - 04

适于数据流组合分类的直推学习方法

刁树民,王永利
(佳木斯大学 公共计算机教研部,黑龙江 佳木斯 154007)
(diaoshumin@126.com)

摘　要:在进行组合决策时,已有的组合分类方法需要对多个组合分类器均有效的公共已知标签训练样本。为了解决在没有已知标签样本的情况下数据流组合分类决策问题,提出一种基于约束学习的数据流组合分类器的融合策略。在判定测试样本上的决策时,根据直推学习理论设计满足每一个局部分类器约束度量的方法,保证了约束的可行性,解决了分布式分类聚集时最大熵的直推扩展问题。测试数据集上的实验证明,与已有的直推学习方法相比,此方法可以获得更好的决策精度,可以应用于数据流组合分类的融合。

关键词:数据流;基于约束学习;直推学习;最大熵;分布式组合分类

中图分类号: TP311.138　　**文献标志码**:A

Transductive learning method applied to ensemble classification over data stream

DIAO Shu-min, WANG Yong-li
(*Department of Common Computer, Jiamusi University, Jiamusi Heilongjiang* 154007, *China*)

Abstract: The existing strategy of combining decisions for ensemble classification method requires common labeled training samples across these ensemble classifiers. To resolve combining classifiers decisions among ensemble classification over data streams without labeled examples, a transductive constraint-based learning strategy was proposed. It satisfied the constraints measured by each local classifier based on transductive learning theory while choosing decision on test samples; thereby guaranteed the feasibility of the constraints. It solved the problems of transductive extension of maximum entropy for aggregation in distributed classification. Experimental examples prove that the proposed method can achieve higher classifying accuracy over the existing transductive approach and can be applied to ensemble classification fusing for data streams.

Key words: data streams; constraint-based learning; transductive learning; maximum entropy; distributed ensemble classification

0　引言

数据流分类是机器学习领域近几年的一个研究热点,组合分类方法以精度高得到了广泛的认可。文献[1]提出了基于多分类器集成学习的概念漂移检测算法 SEA; 文献[2]对多分类器中的权值变化和裁减问题进行了讨论; 文献[3]提出了增量式的 DWM 方法;文献[4]提出 CBEA 方法,集中讨论了一种基于聚类方法的多分类器裁减问题; 文献[5]将流行的 Boosting 技术用于数据流的概念检测中, 提出了自适应多分类器综合挖掘方法;文献[7]提出一种基于多分类器的数据流挖掘算法。上述组合决策系统均是以监督方式设计组合分类器,通过单个分类器的软硬决策获得最后决策,在融合专家知识或学习组合分类器决策的聚集函数时,需要公共的有标记训练数据。如果没有公共有标签数据,就需要采用 Bayes 推理或投票等规则。这种混合规则方法的主要缺点有两个:1)每个单独的分类器可能假定不正确的分类先验概率;2)忽略了单分类器之间的统计依赖关系,或者错误地判断了这种依赖关系。

文献[8]提出一种直推的、基于约束的方法,每个局部分类器贡献的聚集函数必须满足统计约束,通过在测试样本上得出决策,有效地纠正了不正确的先验局部类标签。考虑分类器之间的依赖关系,文献[8]使用启发式梯度下降方法,在测试样本上选择组合后验方法为约束概率编码,这种启发式学习方法存在两个缺陷:由于单个分类器的训练集(用于测度约束的概率)与测试集(用于学习对模型的估计)的不同,不能保证约束的可行性;即使约束是可行的,存在可行方案集,启发式方案也不能确保测试集精确。本文针对分布式分类聚集提出一种最大熵直推式扩展方法,在测试集支持训练的基础上保证约束的可行性,解决了原方法中采用测试集精度偏低的问题。

1　基于约束的数据流组合分类模型

在 Turnstile 模式的基础上[11],定义适合于组合数据流分类的滑动数据流窗口模式。

多维数据流 $\cdots a_i \cdots$ 可定义为多维信号 X 到实数集上的一个映射:$X[1..N] \rightarrow \boldsymbol{R}^p$,每个 a_i 是对 $X[j]$ 更新的值。$a_i = (j, \Delta_i)$ 其含义为 $X_i[j] = X_{i-1}[j] + \Delta_i$,$\Delta_i$ 可能为正也可能为负,表示在时刻 t 的 p 维更新向量,其中每个分量 $\Delta_{i,t}(i = 1, 2, \cdots, p)$ 表示一个属性的更新值。向量 Δ_i 只能读取一次,按照索引(时间戳)i 增加的顺序流入。定义包含最近 n 项元素的序列 $a_{t-n+1}, \cdots, a_t$ 为多维数据流的滑动窗口模式,这样多维数据流可以视为矩阵。这里只是概念意义上的矩阵,并不需要真正物

收稿日期:2008 - 12 - 19;**修回日期**:2009 - 03 - 17。
基金项目:黑龙江省教育厅科学技术项目(11531366), 佳木斯大学科学技术研究项目(L2008-070)。
作者简介:刁树民(1960 -),男,山东龙口人,教授,主要研究方向:数据流挖掘、模式识别;　王永利(1974 -). 男,黑龙江佳木斯人,讲师,博士,主要研究方向:数据流挖掘、数据库。

化这个矩阵。向量 Δ_t 相当于关系表中的元组。

设存在 n 个单独的分类器和 N_c 个类,采用单独的数据流训练集设计第 j 个分类器,训练集样本 $X_j=\{(x_i^{(j)},c_i^{(j)}),i=1,\cdots,N_j\}$,其中 $x_i^{(j)}$ 和 $c_i^{(j)}$ 分别为特征向量和类标签。定义测试样本集 X_{test},分类器 j 根据特征向量 $x_i^{(j)}$ 产生硬、软决策:

$$\{P_j[\hat{C}_j=\hat{c}\mid x^{(j)}]\in[0,1];\hat{c}=1,\cdots,N_c;j=1,\cdots,n\} \tag{1}$$

由聚集函数完成最终决策 $\{P_e[C=c\mid x]\in X_{\text{test}}\}$。

本文采用的组合分类模型,对训练集和测试集采用不同的分类先验概率,不需要公共的有标签训练数据,参照文献[1]将 $\{P_g^{(j)}[\hat{C}_j=\hat{c}\mid C=c],\forall\hat{c},c\}$ 作为局部分类器 j 的约束概率,计算方法为:

$$P_g^{(j)}[\hat{C}_j=\hat{c}\mid C=c]=\frac{\sum\limits_{i=1:c_i^{(j)}=c}^{N_j}P_j[\hat{C}_j=\hat{c}\mid \underline{x}_i^{(j)}]}{\sum\limits_{i=1:c_i^{(j)}=c}^{N_j}1};\forall j,\hat{c},c \tag{2}$$

聚集函数的直推模型估计为:

$$P_m[\hat{C}_j=\hat{c}\mid C=c]=\frac{\sum\limits_{i=1}^{N_{\text{test}}}P_e[C=c\mid \underline{x}_i]P_j[\hat{C}_j=\hat{c}\mid \underline{x}_i^{(j)}]}{\sum\limits_{i=1}^{N_{\text{test}}}P_e[C=c\mid \underline{x}_i]};\forall j,\hat{c},c \tag{3}$$

目标是选取组合后验概率 $P_e[C=c\mid \underline{x}_i]$,以便在与约束匹配的前提下直推估计,即:

$$P_m[\hat{C}_j=\hat{c}\mid C=c]=P_g^{(j)}[\hat{C}_j=\hat{c}\mid C=c];\forall j,\hat{c},c \tag{4}$$

对式(3)最小化得到上面的启发式。

2　最大熵 StreamME 方法

设数据流测试集样本服从均匀分布,即 $P_e[\underline{x}]=1/N_{\text{test}}$,$\forall\ \underline{x}\in X_{\text{test}}$,满足给定的约束(4),采用最大熵基本原理发现均匀条件分布 $P_e[\underline{c\mid x}]$[2,10],由于约束由每个分类器的训练集来度量,所以全局约束可能不可行,本文提出一种 StreamME 方法,尝试采用增强单个测试集的支持来满足约束可行的条件。

2.1 增强单个训练集支持

如果给所有测试集和训练集支持的点分配均匀分布,通常不会确保约束可行性,为保证约束可行性,需要给训练集分配足够的概率密度。因此,我们允许为训练集分配灵活的概率密度(包括为训练集分配的总体密度,训练集中的样本如何分布等)。选择如下形式的联合概率密度函数:

$$P_e[c,\underline{\{x\}}]=\begin{cases}\dfrac{P_u}{N_{\text{test}}}p_e[c\mid \underline{x}], & \underline{x}\in X_{\text{test}}\\ P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}], & \{\underline{x}:\underline{x}^{(j)}=\tilde{\underline{x}}^{(j)}\},\ (\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)})\in X_j\\ 0, & \text{其他}\end{cases} \tag{5}$$

其中:分配给 X_{test} 的总体密度为 $P_u=\sum\limits_{\underline{x}\in X_{\text{test}}}\sum\limits_c p_e[c\mid \underline{x}]$,为每个测试样本分配相同的密度 $\dfrac{P_u}{N_{\text{test}}}P_u$、$p_e[c\mid \underline{x}]$ 和 $P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}]$ 为分配给增强支持点的概率参数。

使用式(5)联合概率密度计算直推估计:

$$P_m[\hat{C}_j=\hat{c}\mid C=c]=\frac{P_m[\hat{C}_j=\hat{c},C=c]}{P_m[C=c]} \tag{6}$$

然而,由于式(5)定义在支持集 $X_{\text{test}}\cup_k X_k$ 上,而后验值 $P_j[\hat{C}_j=\hat{c}\mid \underline{x}^{(j)}]$ 只能在支持子集 $X_{\text{test}}\cup X_j$ 上评价,这意味着不能使用完整的支持集来测度 $P_m[\hat{C}_j=\hat{c},C=c]$。通过调节方法来解决此问题。令:

$$X_r^{(j)}=\{\{\underline{x}\in X_{\text{test}}\}\cup\{\underline{x}:\underline{x}^{(j)}\in X_j\}\} \tag{7}$$

然后我们测度:

$$P_m[\hat{C}_j=\hat{c}\mid C=c,\underline{x}\in X_r^{(j)}]=\frac{P_m[\hat{C}_j=\hat{c},C=c,\underline{x}\in X_r^{(j)}]}{P_m[C=c,\underline{x}\in X_r^{(j)}]} \tag{8}$$

令

$$N_m[\hat{C}_j=\hat{c}\mid C=c,\underline{x}\in X_r^{(j)}]\equiv K_0P_m[\hat{C}_j=\hat{c}\mid C=c,\underline{x}\in X_r^{(j)}] \tag{9}$$

且

$$N_m[C=c,\underline{x}\in X_r^{(j)}]\equiv K_0P_m[C=c,\underline{x}\in X_r^{(j)}] \tag{10}$$

其中 K_0 为两个等式的同一规范化常数。保证这些概率密度函数都为 1,有:

$$P_m[\hat{C}_j=\hat{c}\mid C=c,\underline{x}\in X_r^{(j)}]=\frac{N_m[\hat{C}_j=\hat{c},C=c,\underline{x}\in X_r^{(j)}]}{N_m[C=c,\underline{x}\in X_r^{(j)}]} \tag{11}$$

标记 $N_m[.]$ 反映了事实,这些数量代表了实际发生的联合期望个数,对于 $\underline{x}\in X_r^{(j)}$,此处

$$N_m[C=c,\underline{x}\in X_r^{(j)}]=\frac{P_u}{N_{\text{test}}}\sum_{\underline{x}\in X_{\text{test}}}P_e[c\mid \underline{x}]+\sum_{\tilde{\underline{x}}^{(j)}\in X_j,\tilde{c}^{(j)}=c}P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}] \tag{12}$$

在式(11)~(13)中,约束(4)是需要学习的参数非线性,文献[7]提出减缓约束公式(4)的方法,使其达到线性。假设 $N_m[C=c,\underline{x}\in X_r^{(j)}]>0$,通过 $N_m[C=c,\underline{x}\in X_r^{(j)}]$ 重写等式线性约束:

$$N_m[\hat{C}_j=\hat{c},C=c,\underline{x}\in X_r^{(j)}]=P_g^{(j)}[\hat{C}_j=\hat{c}\mid C=c]N_m[C=c,\underline{x}\in X_r^{(j)}] \tag{13}$$

2.2　直推 StreamME 算法描述

至此 StreamME 算法的形式化表达如下。

参数:

P_u

$\{P[c\mid \underline{x}],\underline{x}\in X_{\text{test}}\}$

$\{P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}],(\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)})\in \tilde{X}_j\}$

最大化:

$$H(c\mid \underline{X})=-\sum_{\underline{x}\in X_{\text{test}}}\frac{P_u}{N_{\text{test}}}\sum_c P_e[c\mid \underline{x}]\log\left(\frac{P_u}{N_{\text{test}}}P_e[c\mid \underline{x}]\right)-\sum_j\sum_{(\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)})\in\tilde{X}_j}P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}]\log P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}]$$

目标:

$$N_m[\hat{C}_j=\hat{c},C=c,\underline{x}\in X_r^{(j)}]=P_g^{(j)}[\hat{C}_j=\hat{c}\mid C=c]N_m[C=c,\underline{x}\in X_r^{(j)}];\forall j,c,\hat{c}$$

$$\sum\nolimits_c P_e[C=c\mid \underline{x}]=1;\forall\ \underline{x}\in X_{\text{test}},$$

$$P_u+\sum_j\sum_{(\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)})\in\tilde{X}_j}P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}]=1$$

$$1-P_u=P_o^*$$

采用Lagrange乘子解决StreamME算法的最大化和目标问题，设$\{\gamma(\hat{C}_j=\hat{c},C=c),\forall j,c,\hat{c}\}$为与单个分类器约束相关联的lagrange乘子，需要学习。自动选择满足约束的汇总，β是外部参数，用于设置先前提到的$1-P_u$。

StreamME算法由优化P_u、$P_e[c\mid \underline{x}]$、$P[\tilde{x}^{(j)},\tilde{c}^{(j)}]$几个步骤组成，在$\{\gamma\}$保持不变和其他参数保持不变时更新$\{\gamma\}$。对于固定的$\beta$，每次迭代时测量偏差的值，即计算$N_m[\hat{C}_j=\hat{c},C=c,x\in X_r^{(j)}]$与$P_g^{(j)}[\hat{C}_j=\hat{c},C=c]N_m[C=c,\underline{x}\in X_r^{(j)}]$的平方Euclidean距离[9]，累加$j,c,\hat{c}$，当$\Delta D$值小于某阈值时停止。StreamME算法递减$L(\beta)$的值，在此解决一个凸优化问题，在满足每个$\beta$约束的情况下，求解唯一的最大熵解，这一方案是可行的。如果对于给定的β问题不可解，就不能保证算法的收敛性。当$\beta\to\infty$、$P_u=1$时，解仅依赖于测试集支持。当$\beta\to\infty$、$P_u=0$时，根本就不必使用测试集。当$\beta=0$，在P_u上没有约束，与在其他β值求得的解相比，此解得到了最高的熵值$H(C,\underline{X})$。

算法 StreamME

输入：滑动窗口样本矩阵$\boldsymbol{X}$，P_u，$\{P[c\mid \underline{x}],\underline{x}\in X_{\mathrm{test}}\}$，$\{P[\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)}],(\tilde{\underline{x}}^{(j)},\tilde{c}^{(j)})\in\tilde{X}_j\}$，$\beta$，$\delta$，当前时间为$\tau$

输出：最大熵解$H(C,\underline{X})$

```
1)  while T(i,t) arrived, for (t ≥ τ - n) do begin
2)      NM = N_m[Ĉ_j = ĉ, C = c, x ∈ X_r^(j)];
3)      PGJ = P_g^(j)[Ĉ_j = ĉ, C = c] N_m[C = c, x ∈ X_r^(j)];
4)      ΔD = NM 与 PGJ 的平方 Euclidean 距离
5)      if (t_1 = t - n) and (ΔD < δ) begin Stop;
                    /* 样本已经过期或 ΔD 超出阈值则停止 */
6)      else
7)        优化 P_u、P_e[c | x]、P[x̃^(j), c̃^(j)];
8)        H(C,X) = 求解唯一的最大熵解;
9)        t_1 = t，累加 j,c,ĉ，递减 L(β) 的值;
                    /* 设置当前时刻为新的求解时刻 */
10)     endif
11)   enddo
```

对β应用扩展二分查找方法查找P_o^*。总体算法在两个条件之一得到满足时停止：

1）在当前值β，未达到约束满足，表示P_o^*大于值$1-P_u$。

2）对于当前值β，StreamME的结束处，$1-P_u<\delta$，其中δ是一个很小的数。仅使用测试支持，存在一个最大熵解满足约束，即$P_o^*\approx 0$。算法复杂度为$O(n^2\mathrm{lb}\,n)$，证明略。

3 性能测试

使用同样在文献[6]中用到的UC Irvine机器学习数据集传感器（Dodgers loop sensor包含50400个样本，3个属性，分属6个事件类别、车辆测试集（vehicle silhouettes，包含946个样本，18个属性，分属4个类别），测试算法的性能。实验计算机的配置是3.0 GHz Pentium/1 GB/200 GB，OS为Windows XP，以Matlab与VC++ 6.0为工具实现算法。算法可以以一次查询和连续查询方式运行，模拟分布式数据流发生器，使用四个不同的分类器（naïve bayes分类器）。因数据集的所有特征子集是连续值，可建立高斯密度模型，在四次执行中对所有的数据集进行双重交叉验证，测量十次测试循环的平均误差，评价测试数据的先验分类精度变化。通过对原始测试集副本的采样得到带有指定先验的新测试集，量化测试集先验（$P_{\mathrm{test}}[C=c]$）和训练集先验（$P_{trn}^{(j)}[C=c]$）。

文献[1]给出了仅使用$\{\{P_e[c\mid \underline{x}]\},\underline{x}\in X_{\mathrm{test}}\}$约束不可满足的样例，即启发式方法不能满足所有的约束，如果不使用支持增强，这个结论对于StreamME算法仍然是成立的。图1、2显示了带有和没有训练支持增强的StreamME算法的Lagrangian代价函数变化。由图1可以看出Lagrangian代价函数单调下降且未收敛，而偏差没有到零，表示此解不是可行解。图2表示具有单独训练支持增强的StreamME算法（$\beta=0$）可以获得可行解。

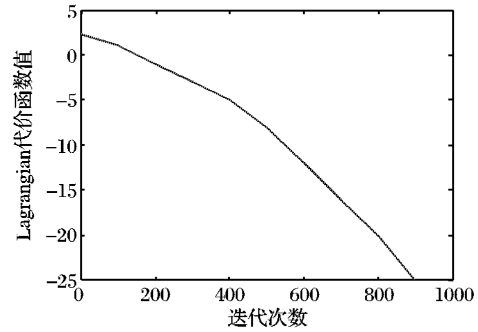

图1 没有训练支持增强的StreamME算法Lagrangian代价函数

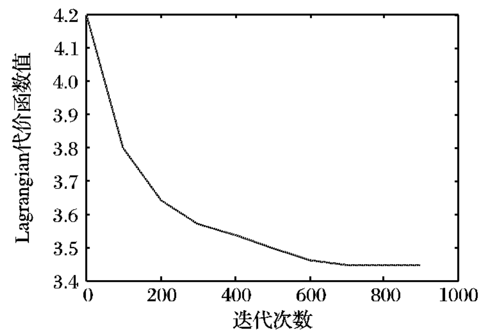

图2 有训练支持增强的StreamME算法Lagrangian代价函数

图3、4显示了在传感器数据集上β对StreamME算法性能的影响，测量联合熵、测试数据上的分类误差率，记录了三次实验的打点曲线，其中$M=\sum_{j=1}^{M_e}\sum_{c=1}^{N_c}P_{trn}^{(j)}[C=c]\times\log\left(\frac{P_{trn}^{(j)}[C=c]}{P_{\mathrm{test}}[C=c]}\right)$，以$\Delta\beta=1$的步长从$-20$到10改变$\beta$的值，对于每个$\beta$，运行StreamME直到$\Delta D<10^{-9}$。

在图3中，当$\beta<-6$时，在StreamME终止时偏差无法抵达0。这说明在限定的β可能产生不可行的约束（当P_l较小时）。StreamME在$\beta=0$时达到峰值熵，如图4所示（密度函数上没有约束），当β为负或正时联合熵降低。P_u的增长率随着β变负逐渐降低，因为如果StreamME算法为测试集分配了过多的概率密度，将很难满足局部约束。

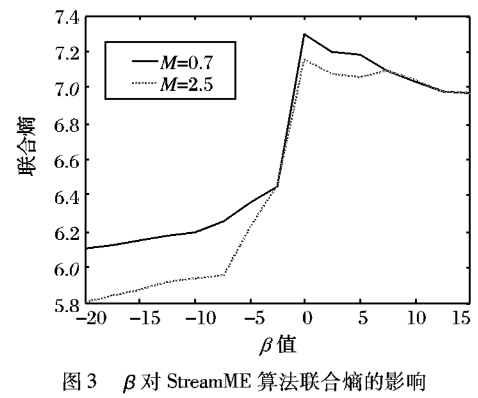

图3 β对StreamME算法联合熵的影响

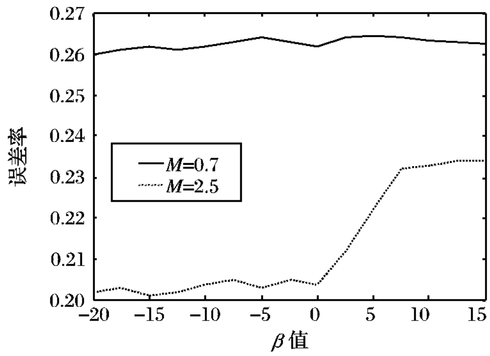

图 4　β 对 StreamME 算法分类误差的影响

采用传感器数据集和车辆数据集对 StreamME 算法和 CB 算法进行了对比实验,表 1 给出了 StreamME 方法与 CB 方法在分类性能方面的比较,误差率 $\bar{P}$ 定义为:$\Delta p_i = \frac{(v_i^c - v_i^a)}{v_i^a} \times 100\%$,其中 v_i^c 表示量化的预测值,v_i^a 表示量化的实际值,$\bar{P} = \left[1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta p_i^2}\right] \times 100\%$,物理含义为平均预测精度。

表 1　StreamME 方法与启发式 CB 方法的分类性能比较

数据集	M	StreamME		heuristic CB	
		误差率/%	条件熵	误差率/%	条件熵
传感器	0.95	52	0.87	58	0.09
	2.70	50	0.76	60	0.11
	missing	48	0.73	62	0.08
车辆	0.89	27	0.34	32	1E−3
	2.28	19	0.25	34	1E−6
	missing	17	0.21	34	1E−6

从 StreamME 与启发式 CB 方法的分类性能比较表可以看出,由于 StreamME 采用了寻找最大熵解的思想,StreamME 可以获得更好的分类误差率,得到期望中更高的熵。

4　结语

本文提出一种新颖的适用于数据流组合分类的直推学习方法,在多个分类器之间没有公共标签数据的情况下,如何选择决策聚集规则。本方法解决了约束不可行、解不唯一等问题,在训练集上测算约束概率,通过对训练集的增强保证了约束的可行。实验证明,在给定测度约束前提下,本文提出的 StreamME 算法的精度比基于约束的启发式方法要高,获得了较好的性能,可广泛应用于传感器数据分类等领域。

参考文献:

[1]　WANG H, FAN W, YU P S. Mining concept-drifting data streams using ensemble classifiers [C]// Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 226 −235.

[2]　KOLTER J Z, MALOOF M A. Dynamic weighted majority: a new ensemble method for tracking concept drift [C]// Proceedings of the 3rd IEEE Conference on Data Mining. Los Alamitos, USA: IEEE Computer Society Press, 2003: 123 −130.

[3]　KOLTER J Z, MALOOF M A. Using additive expert ensembles to cope with concept drift [C]// Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany: ACM Press, 2005: 449 −456.

[4]　RUSHING J, GRAVES S, CRISWELL E, *et al.* coverage based ensemble algorithm (CBEA) for streaming data [C]// Proceedings of the 16th IEEE International Conference on Tools with Artifficial Intelligence. Huntsville, USA: IEEE Computer Society Press, 2004: 106 −112.

[5]　CHU F, ZANIOLO C. Fast and light boosting for adaptive mining of data streams [C]// Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Sydney, Australia: Springer Press, 2004: 282 −292.

[6]　MILLER D J, PAL S. Transductive methods for the distributed ensemble classification problem [J]. Neural Computation, 2007, 19 (3): 856 −884.

[7]　孙岳, 毛国君, 刘旭, 等. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报, 2008, 34(1): 94 −96.

[8]　KRISHNAPURAM B, WILLIAMS D, XUE YA, *et al.* On semi-supervised classification [EB/OL]. [2008 −10 −10]. http://www.lx.it.pt/~mtf/NIPS_2005.pdf.

[9]　JAYNES E T. Papers on probability, statistics, and statistical physics [M]. Dordrecht, Holand: Kluwer Academic Publishers, 1989.

[10]　PIETRA S D, BERGER A L, PIETRA V J. A maximum entropy approach to natural language processing [J]. Computational Linguistics, 1996, 22(1): 39 −71.

[11]　MUTHUKRISHNAN S. Data streams: Algorithms and applications [EB/OL]. [2008 −10 −10]. http://www.cs.rutgers.edu/~muthu/stream-1-1.ps.

(上接第 1562 页)

类熵精度 β_a 和 β_b,对 a 类与 b 类训练样本集分别进行合理的聚类,甚至可以保证在 b 类识别率基本保持不变的情况下,a 类的识别率有所提高,如表 2 中序号为 3 和 4 的数据所示。

通过对实验结果的分析,可以发现本文所提出的新的稀疏化算法确实具有很大的优势。

6　结语

本文提出了一种新的最小二乘支持向量机的稀疏化算法,并对其进行了详尽的阐述。本文所提出的稀疏化算法完全可以应用到计算机或雷达图像处理、DNA 检测、故障诊断以及其他模式识别等领域。

参考文献:

[1]　陈爱军. 最小二乘支持向量机及其在工业过程建模中的应用[D]. 杭州: 浙江大学, 2006.

[2]　朱家元, 陈开陶, 张恒喜. 最小二乘支持向量机算法研究[J]. 计算机科学, 2003, 30(7): 157 −159.

[3]　边肇祺, 张学工. 模式识别[M]. 2 版. 北京: 清华大学出版社, 2000: 296 −301.

[4]　田盛丰. 基于核函数的学习算法[J]. 北方交通大学学报, 2003, 27(2): 1 −8.

[5]　RENYI A. On measures of entropy and information [EB/OL]. [2008 −10 −10]. http://digitalassets.lib.berkeley.edu/math/ucb/text/math_s4_v1_article-27.pdf.

[6]　GIROLAMI M. Orthogonal series density estimation and the kernel eigenvalue problem [J]. Neural Computation, 2002, 14(3): 669 −688.

[7]　吴宗亮, 窦衡. 一种广义最小二乘支持向量机算法及其应用[J]. 计算机应用 2009, 29(3): 877 −879.