最小最大模块化支持向量机改进研究

文益民[1,2]　吕宝粮[1]
[1](上海交通大学计算机科学与工程系,上海 200030)
[2](湖南工业职业技术学院,长沙 410007)
E-mail:ymwen@cs.sjtu.edu.cn

摘　要　该文提出了一种新的聚类算法以实现训练数据的等分割并将其应用于最小最大模块化支持向量机(M^3-SVM)。仿真实验表明:当训练数据不是同分布时,与随机分割方法相比,该文提出的聚类算法不但能提高 M^3-SVM 的一般化能力,缩短训练时间,还能减少支持向量。

关键词　等分聚类　支持向量机　最小最大模块化支持向量机

文章编号 1002-8331-(2005)19-0185-04　**文献标识码** A　**中图分类号** TP311.13

Improvement Research of Min-Max Modular Support Vector Machine

Wen Yimin[1,2]　Lv Baoliang[1]
[1](Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030)
[2](Hunan Industry Polytechnic, Changsha 410007)

Abstract: In this paper, a new cluster method that can equally partition training data is proposed for min-max modular support vector machine (M^3-SVM). The experiments illustrate that the proposed cluster method can not only increase the generalization accuracy of M^3-SVM, but also speed up training and reduce the number of support vectors in comparison with the existing random data decomposition method.

Keywords: equal clustering, Support Vector Machine, Min-Max Modular Support Vector Machine

1　引言

由于信息技术的飞速发展,以及系统决策的要求,在很多领域产生了海量数据的机器学习问题,这些领域包括气象数据、Web 数据、商务数据和国民经济统计数据的处理等。现有的机器学习算法如:神经网络、贝叶斯分类器、决策树、支持向量机等在海量数据的处理问题上都存在训练时间过长的问题。如何将现有的各种机器学习算法扩展到海量数据的处理问题上是机器学习的研究热点之一。Provost 指出并行化是扩展现有机器学习方法的一条重要途径[1]。在训练阶段,并行化机器学习方法通常将训练集分解成若干规模适度的子集,将各子集并行地使用传统的机器学习方法进行处理,得到各自的子学习器。在测试阶段,将待测试样本提交给各个子学习器,由各个子学习器给出它们各自的预测结果,然后采用某种集成方法将各个子学习器的结果集成后得到最终结果。显然这种并行化机器学习方法的性能决定于训练集的分解方法和各子学习器结果的集成规则。

文献[2]提出用最小最大模块化(Min-Max Modular, M^3)方法来扩展现有机器学习方法。通过采用随机分割训练集的方法和"最小"和"最大"两条集成规则,M^3 已经成功地扩展了神经网络方法(M^3-MLP),将其应用于脑电信号分类[3]。M^3 还成功地扩展了支持向量机方法(M^3-SVM),将其应用于文本分类[4]。从本质上讲,M^3 的两条集成规则能根据测试样本动态地选择一个子分类器对测试样本给出判断。因此,应根据数据的分布特征对训练集进行分解。当训练数据不是同分布时,尤其需要这样。随机分割训练集的方法忽视了数据本身的分布特征,给 M^3 学习器的一般化能力带来负面影响。该文提出了一种新的等分聚类分割算法。基于这种新的数据分割算法,与随机分割相比,M^3-SVM 的一般化能力得到了提高,但训练时间却缩短,同时支持向量的数目减少。

2　最小最大模块化支持向量机

给定一个两类分类问题 $S=X^+\cup X^-$,其中 $X^+=\{(X_i,+1)\}_{i=1}^{N^+}$ 表示正类样本集,$X^-=\{(X_i,-1)\}_{i=1}^{N^-}$ 表示负类样本集。其中:X_i 表示第 i 个训练样本,N^+ 和 N^- 分别表示正类和负类样本的数目,则训练样本总数为:$N=N^++N^-$。

在训练阶段:根据事先确定的分解常数 K^+ 和 K^-,按照某种训练集分解方法,将原训练集 X^+ 和 X^- 分别分解为 K^+ 和 K^- 个包含样本数量大致相等且互不相交的子集:

$$X^+=\bigcup_{i=1}^{K^+}X_i^+,\ \bigcap_{i=1}^{K^+}X_i^+=\Phi,X_i^+=\{(X_m,+1)\}_{m=1}^{N_i^+},i=1,2,\ldots,K^+$$
$$X^-=\bigcup_{j=1}^{K^-}X_j^-,\ \bigcap_{j=1}^{K^-}X_j^-=\Phi,X_j^-=\{(X_n,-1)\}_{n=1}^{N_j^-},j=1,2,\ldots,K^- \quad (1)$$

其中:$\sum_{i=1}^{K^+}N_i^+=N^+$,$\sum_{j=1}^{K^-}X_j^-=N^-$;$\Phi$ 表示空集。于是原两类分类

基金项目:国家自然科学基金(编号:60375022,60473040)资助
作者简介:文益民(1969-),男,湖南桃江人,博士生,主要研究领域为统计学习理论,生物信息学,图像处理。吕宝粮(1960-),男,工学博士,教授,博士生导师,IEEE 高级会员,主要研究领域为仿脑计算机理论与模型,神经网络,机器学习,计算系统生物学,自然语言处理。

问题 S 被分解成下列 $K^+\times K^-$ 个规模较小的两类分类子问题：

$$S_{i,j}=X_i^+\cup X_j^-,i=1,2,\dots,K^+,j=1,2,\dots,K^- \quad (2)$$

由于这 $K^+\times K^-$ 个子问题在处理时不需要相互通信，可将得到的这 $K^+\times K^-$ 个子问题 $S_{i,j}$ ($i=1,2,\dots,K^+,j=1,2,\dots,K^-$) 使用通常的支持向量机(SVM)训练方法并行或串行训练，得到 $K^+\times K^-$ 个子分类器，分别表示为：$SVM_{i,j}$ ($i=1,2,\dots,K^+,j=1,2,\dots,K^-$)。训练集需要等分是为了保证并行训练各子分类器时各处理器的负载平衡，但需要指出的是：支持向量机的训练时间并不唯一地决定于训练集的大小。

在测试阶段，将测试样本 X 提交给所有的子分类器 $SVM_{i,j}$ ($i=1,2,\dots,K^+,j=1,2,\dots,K^-$)，各分类器的输出为：$SVM_{i,j}(X)$。通过“最小”原则，将得到：

$$G_i(X)=\max_{j=1}^{K^-} SVM_{i,j}(X),i=1,2,\dots,K^+ \quad (3)$$

然后通过“最大”原则，得到：

$$C(X)=\max_{i=1}^{K^+} G_i(X) \quad (4)$$

最后可以根据 $C(X)$ 的值对 X 的类别做出判断。

3 基于等分聚类的数据分割算法

训练数据分割的问题可以描述为：设有样本集 $S=\{X_i|X_i\in R^n,i=1,2,\dots,N\}$，要将 S 等分割成 K 个在空间上互不重叠的子集合：$S_1,S_2,\cdots,S_K$。这个划分过程的数学模型可以描述为以下的非线性规划问题：

$$Minimize_{s_1,s_2,\dots,s_K}: Dif=\max_{i=1}^{K}\left|W_i-\overline{W}\right|$$

式中：s_i 和 W_i 分别表示各个子集合的中心向量和各个子集合中的样本数量。K 表示需要划分的子集合数目。$\overline{W}$ 表示各子集合的平均样本数目。显然有，$\overline{W}=\lfloor\frac{N}{K}\rfloor$。目标函数值 Dif 用来衡量各个子集合中样本分布的平衡程度，Dif 的值越小，各个子集合中的样本数量越均匀，此时的划分效果就越好。

提出的等分聚类分割算法如下：

算法输入：待划分样本集 S；预先设定的划分子集合数 K；最大迭代步数 max$iter$；算法参数 α 和 l；目标函数的阈值 ε。

初始化：随机选择 S 中的 K 个样本作为子集的中心 s_i^0，$i=1,2,\dots,K$。

算法过程：

For $t=1$：max$iter$

1.根据 S 中各样本与各子集合中心的距离将各个样本划分给各个子集合，记录各个样本所属的子集合的编号：

$$subclusterlabel[j]=\arg\min_{i=1}^{K}\|X-s_i^{t-1}\|,j=1,2,\cdots,N$$

2.计算 $W_i(i=1,2,\dots,K)$ 的大小

3.计算目标函数 Dif 的值

4.如果：$Dif<\varepsilon$，就中断循环

5.For $i=1:K$

5.1 计算 $\delta v_i=\sum_{j=1,j\neq i}^{K}(\frac{l*W_j}{W_j+(l-1)*W_i}-1)(s_j^{t-1}-s_i^{t-1})$

5.2 更新子集合的中心向量：$s_i^t=s_i^{t-1}+\alpha*\delta v_i$

以上算法的工作原理是：如果两个子集合包含的样本数量不相等，则包含样本数量较少的子集合 S_i 的中心 s_i 会向包含样本数量较多的子集合 S_j 的中心 s_j 靠近。而 s_j 也沿着向量 s_j-s_i 的方向移动一定的距离，也就是 s_i 和 s_j 同向移动。这样 S_i 中的样本会增加，而 S_j 中的样本会减少。显然该算法的时间复杂度不会超过 $O(K*\max iter)$。该聚类算法只要一次顺序扫描外存储器就可以完成一次迭代，因此空间复杂度为 $O(1)$。因此不管待划分的集合有多大，该文提出的聚类算法都能实现样本集合的等分。因此这种聚类分割算法用于海量数据的处理是合适的。

该文提出的算法与文献[5]中提出的算法相比作了两个重要的改进：其一，在算法中加入了算法中断条件：如果 $Dif<\varepsilon$，就中断迭代过程。这个改进能保证算法一旦满足了预先设定的平衡分布条件，就结束算法。如果没有这个条件，当达到预先设定的迭代次数 max$iter$ 时，划分的效果可能不会太好。这是因为 Dif 在迭代过程中不完全是单调下降的。其二，在子集合的中心修改方法中将原来的 $\frac{W_j}{W_i}$ 改为现在的 $\frac{l*W_j}{W_j-(l-1)*W_i}$，这避免了当两个子集合中样本数量相差很大时 $\frac{W_j}{W_i}$ 的值过大的问题。因为 $\frac{W_j}{W_i}$ 的值过大会使得新的子集合中心与原中心位置相差过大，不利于各子集合中样本数量的均匀分布。这将导致算法不能收敛，并在样本密集且不同分布的数据集中尤为突出。为了方便起见称文献[5]提出的算法为 GeoClust，该文提出的算法为 GeoClustM。

4 仿真实验与讨论

在该文的仿真实验中，使用的数据及相关参数如表 1 所示。Banana 数据来自文献[6]，Letter recognition 和 Forest Cover Type 数据来自文献[7]。其中 Letter recognition 和 Forest Cover Type 两个数据集中的样本的各特征均规范至[0,1]。Letter recognition 原本是一个 26 类问题，采用文献[8]提出的方法将其变为两类问题。将“E”、“H”、“I”、“M”、“P”、“Q”、“R”、“X”、“Y”和“Z”这些类别的样本看成正类样本，其余类别看成负类样本。使用其中 $\frac{3}{4}$ 的数据作为训练数据，剩余 $\frac{1}{4}$ 作为测试数据。在 Forest CoverType 中将第二类样本看成是正类样本，而将所有其他 6 类样本看成是负类样本。将其中一半用于训练，剩余一半用于测试。聚类分割算法各参数的设置为：max$iter$=6 000，$\alpha=0.01*10^{-\lfloor\frac{K-1}{10}\rfloor}$，$l=3$，$\varepsilon=\lfloor\frac{N}{50K}\rfloor$。由于很难知道最佳的划分数，划分数 K 从 2 取到 20。所有的实验均重复 3 次，最后取它们的平均值。在该文中，仿真实验平台为 1GRAM3GCPU 的 PC。

表 1 各实验数据集及相关参数

问题	样本特征数	类别数	训练数据	测试数据	c	σ
Banana	2	2	40 000	49 000	316.2	0.707
Letter recognition	16	2	15 000	5 000	16	4
Forest CoverType	54	2	290 504	290 508	128	0.25

为了比较 GeoClust 和 GeoClustM 的性能，使用了数据密集且不同分布的 Banana 数据集中的数据。为了综合考虑算法的划分平衡程度 Dif 与迭代次数 $iternum$ 的关系，定义了变数量：$\mu=Dif\times iternum$。显然，μ 值越小算法的划分性能越佳。从图 1 可以看出当数据密集且不同分布时，GeoClust 更容易使得子集合

表 2　M³–SVM–RP 和 M³–SVM–CE 的一般化能力比较

		K=2,3,4,…,20 时 M³–SVM 的一般化能力(1/10000)																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
B	RP	**9084**	**9084**	9081	9090	9083	9085	9081	9088	9094	9087	9085	**9079**	9092	9091	9075	9081	9083	9089	9088
	CE	9081	9080	9094	9091	9093	9108	9110	9103	9100	9093	9118	9073	9101	9105	9101	9107	9099	9103	9093
L	RP	9899	9899	9895	9896	9907	**9920**	9918	9913	9912	9919	9915	9918	9929	9925	9917	9933	9933	9919	9927
	CE	9910	9929	9923	9926	9911	9919	9935	9939	9946	9936	9927	9939	9942	9943	9937	9953	9941	9953	9955
F	RP	8742	**8681**	8626	8604	8574	8539	8530	8510	8481	5452	8452	8440	8425	8413	8402	8396	8390	8385	8376
	CE	8747	8659	8762	8778	8762	8652	8760	8689	8747	8704	8689	8768	8689	8792	8823	8807	8807	8763	8708

表中第一列中“B”、“L”和“F”分别表示 Banana、Letter recognition 和 Forest CoverType 数据集。黑体字表示此时 M³–SVM–RP 的一般化能力要比对应 M³–SVM–CE 的一般化能力要好

的中心移动的距离过大，使得算法不容易收敛。图 2 则表明：在 Banana 数据集上 GeoClust 划分性能比 GeoClustM 要差。

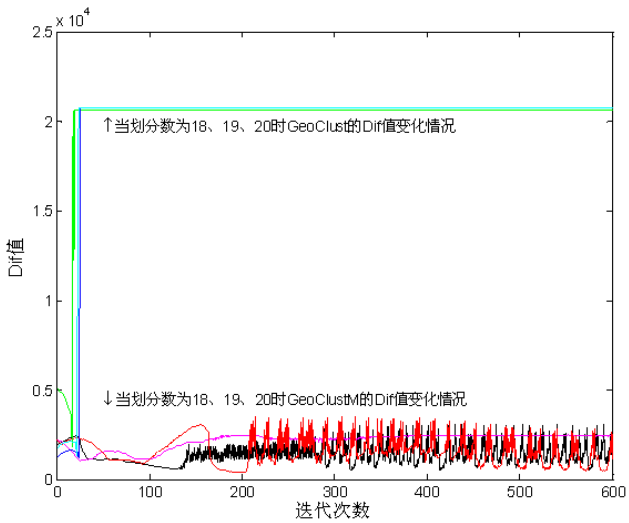

图 1　GenClustM 在 Banana 收敛性能比较

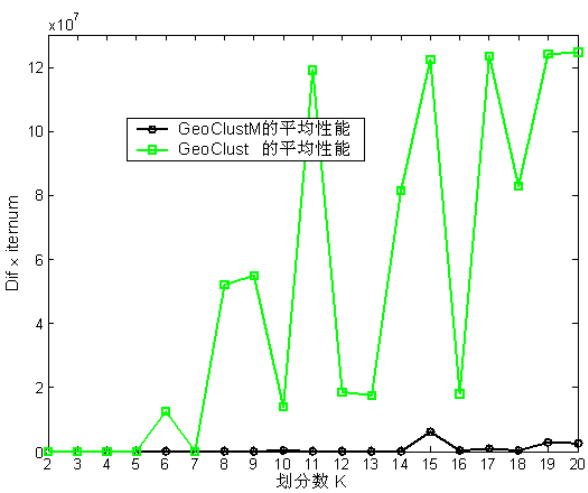

图 2　GenClustM 与 GenClust 的划分性能比较

为了比较等分聚类分割与随机分割对 M³–SVM 性能的影响，使用 libSVM 作为训练子分类器的工具，libSVM 的 cache 被设置为 100M。采用的核函数为高斯核函数：$\exp(-\frac{1}{2\sigma^2}\|X-X_i\|^2)$。为了简单起见，定义 $K^+=K^-$，将它们统称为 K。采用两种方式统计 M³–SVM 的训练时间：一种是串行时间，也就是所有子分类器的训练时间之和；另一种是并行时间，也就是所有子分类器的训练时间中的最长时间。

在该文各图表中，“RP”表示随机分割算法，“CE”表示等分聚类分割算法。M³–SVM–RP 表示基于随机分割的最小最大模块化支持向量机方法。M³–SVM–CE 的意义可以类推。

表 2 说明在绝大多数情况下 M³–SVM–CE 的一般化能力要比 M³–SVM–RP 好。而图 3、图 4 和图 5 则说明 M³–SVM–CE 的支持向量数目比 M³–SVM–RP 的支持向量数目要少很多。这意味着在进行测试时 M³–SVM–CE 将比 M³–SVM–RP 快。在图 6、图 7 和图 8 中，即使考虑到表 3 中聚类分割的时间消耗，三个实验中 M³–SVM–CE 串行训练时间都要比 M³–SVM–RP 的串行训练时间少得多。

表 3　RP 和 CE 对训练集进行一次 K 等分划分的平均时间

数据集	Banana		Letter recognition		Forest CoverType	
数据集划分方法	RP	CE	RP	CE	RP	CE
一次 K 等分划分的平均时间/s	0.6	90.1	1.5	10.6	19.6	756.3

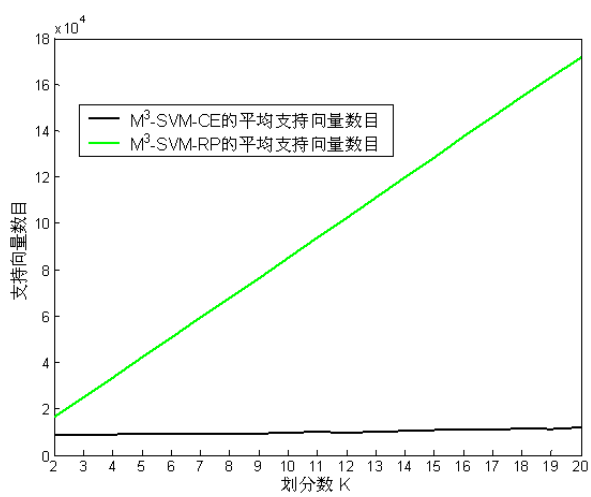

图 3　M³–SVM 在 Banana 上的支持向量数目增长

三个实验中当划分数 K 较小时，即使考虑聚类分割的时间消耗，M³–SVM–CE 的并行训练时间都要比 M³–SVM–RP 的并行训练时间少。当划分数 K 较大时，由于每个子问题的规模减小，各分类器的训练时间将减少。此时如果考虑聚类分割的时间消耗，M³–SVM–CE 的并行训练时间则会比 M³–SVM–RP 的并行训练时间多，但是支持向量数目会更少。如果完全不计训练集的划分时间，则当划分数 K 为任意值时，M³–SVM–CE 的并行训练时间都要比 M³–SVM–RP 的并行训练时间少。以上总体说明：M³–SVM–CE 的性能要超过 M³–SVM–RP。

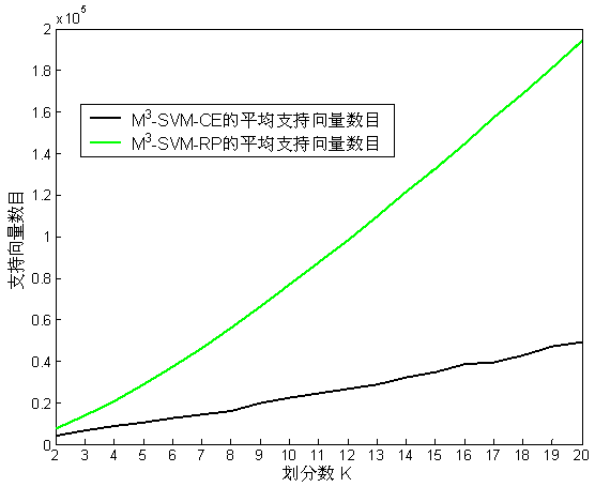

图 4　M3–SVM 在 Letter 上的支持向量数目增长

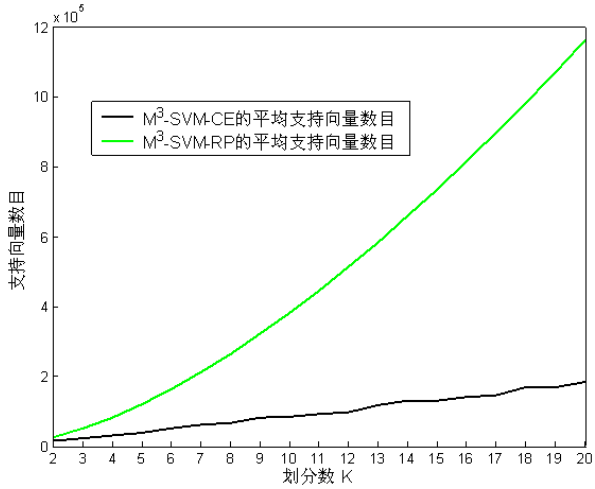

图 5　M³–SVM 在 Forest CoverType 上的支持向量数目增长

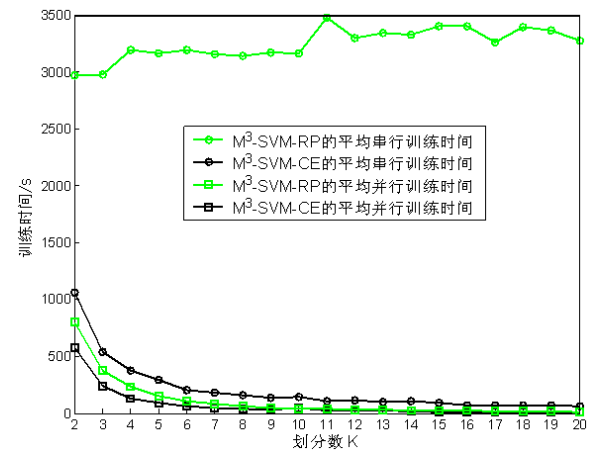

图 6　M³–SVM 在 Banana 上的训练时间比较

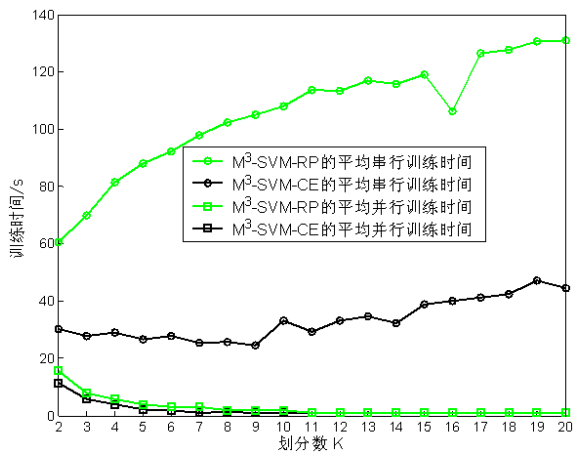

图 7　M³–SVM在 Letter 上的训练时间比较

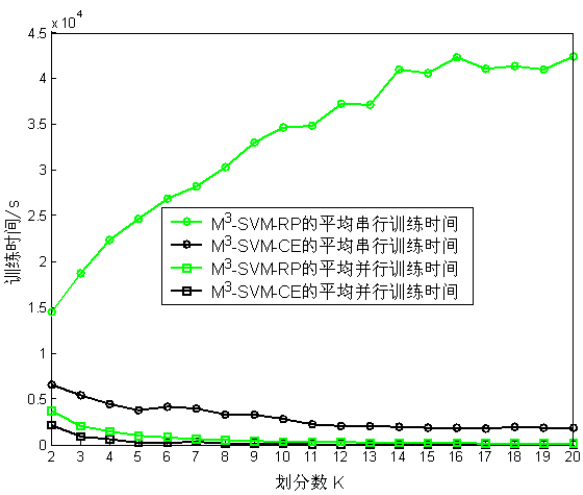

图 8　M³–SVM 在 Forest CoverType 上的训练时间比较

等分聚类分割能使 M³–SVM 性能提高的原因在于：随机划分训练集后，每个子分类问题中数据的分布将与原问题中数据的分布相似，训练后得到的各个子分类器也将相似。这会使得 M³–SVM 的两条集成规则的动态挑选子分类器的作用不能得到充分发挥，这将影响 M³–SVM 的一般化能力。与此相反，当训练数据不是同分布时，聚类划分训练集后，各子分类器由位于特征空间不同位置且数据分布完全不同的子样本集训练得到，这将充分发挥 M³–SVM 的两条集成规则的动态挑选子分类器的作用，因此能提高 M³–SVM 的一般化能力。另外，聚类划分训练集后，与随机划分相比，各子分类问题中的正负两类样本的可分性增大，因此 M³–SVM–CE 的支持向量要比 M³–SVM–RP 少，训练时间也因此而减少。根据 Vapnik(1995)的观点，支持向量的减少通常会导致支持向量机一般化能力的提高，因此最终提高了 M³–SVM 的一般化能力。

5　结论

该文提出了一种新的聚类算法实现对训练集的等分割，它的特点是能比较均匀地实现训练集的分割，尽可能保证 M³–SVM 在并行执行时各处理器的负载平衡。仿真实验表明：当训练数据分布不是同分布时，聚类分割方法使数据的分割体现数据本身的分布特征，能尽量减少因数据分割而带来的分类信息损失，而随机分割显然忽略了这一点。该文提出的等分聚类分割算法能显著提高 M3–SVM 的性能。(收稿日期：2005 年 5 月)

参考文献

1.Provost F，Aronis J M.Scaling up inductive learning with massive par–

(下转 198 页)

$$\begin{cases} D=23n_1\pm3 & \text{体力节律协调} \\ D=28n_2\pm4 & \text{情绪节律协调} \\ D=33n_3\pm5 & \text{智力节律协调} \end{cases}$$

5 研究与讨论

为进一步验证该系统对人体生物节律,特别是人体生物节律优生法则的实行情况,笔者使用该系统展开了多项调查,下面列举其中具有代表性的两个。

对2004年获得雅典奥运会单人和双人项目金牌的40名中国运动员进行了分析,统计了他们在获得金牌的日期以及获金牌日期的前2天的三节律所处的阶段。可以发现,体力因素对运动员获得好成绩的影响最大,情绪次之。因此在同等实力水平的运动员中挑选出赛者时可以以此为参考。如表2所示。

表2 中国奥运金牌得主在比赛日的三节律分布

	体力	情绪	智力
高潮期60~100	19人	16人	14人
临界期-60~60	13人	12人	13人
低潮期-100~-60	8人	12人	12人

对2001~2003年在妇幼保健院出生的200名婴儿以及他们的父母作了调查。根据世界卫生组织的标准对婴儿的身体发育和疾病发生情况进行评估,并结合孩子和父母的生日以及母亲的末次月经计算出父母在受孕时的平均体力节律值,这些数据一定程度上说明了符合体力优生的婴儿,先天的体质较好。如表3所示。

表3 婴儿常见病患病率的比较

受孕时父母的平均体力节律值	发病人数	调查人数	异常率/%
高潮期60~100	34	65	52.30
临界期-60~60	61	87	70.14
低潮期-100~-60	38	48	79.16

6 总结

人体生物节律优生辅助系统可以方便、有效地帮助用户统计和比较人体生物节律,为优生优育提供了有效的分析和决策工具,为孕、产期的科技智能化管理提供了新的途径,可以进一步堵截或减少出生缺陷儿的发生,提高人口素质。笔者还将对该系统进行改进,让它在更多的领域得到应用。

(收稿日期:2005年2月)

参考文献

1.Dr Donald,L McEachron.Time and time again-The effects of biological rhythms on human health and performance[C].In:Frequency Control Symposium and Exhibition 2000,Proceedings of the 2000 IEEE/EIA International,2000-06:7~21

2.H Hagiwara,T Nakano,K Yoshida et al.Measurement of human behavior in a daily life based on the understanding of biological rhythm[C].In:Proceedings of the 41st SICE Annual Conference,2002;2:789~793

3.李晓明.试溯人体生物节律之源[J].科学技术与辨证法,1995;12(2):22~25

4.朱静华.人体"生物三节律"在田径训练和比赛中的运用[J].首都体育学院学报,2002;14(2):66~69

5.马丽华.应用人体生物节律作好事故预防[J].水利电力劳动保护,2003;4:18~19

(上接188页)

allelism[J].Machine Learning,1996;23:33~46

2.Lu B L,Ito M.Task decomposition and module combination based on class relations:a modular neural networks for pattern classification[J].IEEE Trans on Neural Networks,1999;10:1244~1256

3.Lu B L,Shin J,Ichikawa M.Massively parallel classification of single-trial EEG signal using min-max modular neural network[J].IEEE Trans on Biomedical Engineering,2004;51:551~558

4.Lu B L,Wang K A,Utiyama M et al.A part-versus-part method for massively parallel training of support vector machines[C].In:Proceedings of IJCNN'04,2004:735~740

5.Choudhury A,Nair C P,Keane A J.A data parallel approach for large-scale gaussian process modeling[C].In:Proceedings of the second SIAM International Conference on Data Mining,2002

6.Rätsch G.http://ida.first.gmd.de/raetsch/data/benchmarks.htm

7.Blake C L,Merz C J.UCI.ftp://ftp.ics.uci.edu/pub/machine-learning-database,1998

8.Rätsch G,Onoda T,Müller K R.Soft margins for adaboost[J].Machine Learning,2001;42:287~320

(上接192页)

目前,基于GSPN的性能评价分析方法还处在完善发展阶段,基于这些方法和技术的完善的自动化建模分析软件工具还很少,大部分的工作还需要以手工方式完成。因此,无论是在基础理论方面,还是在工程应用方面,如何更加有效地使用GSPN对工作流模型进行性能评价分析,值得进一步的研究。

(收稿日期:2005年4月)

参考文献

1.Workflow Management Coalition.The Workflow Reference Model.WFMC TC00-1003,1995-01

2.林闯.计算机网络和计算机系统的性能评价[M].清华大学出版社,2001

3.Zuberek W K.Performance Evaluation Using Unbound Timed Petri Nets[C].In:Proc of the Third International Workshop on Petri Nets and Performance Models,Kyoto,Japan,1989:180~186

4.Dugan J B,Trivedi K S,Geist R M et al.Extended stochastic Petri nets:Applications and Analysis[C].In:Gelenbe E ed.PERFORMANCE84,Mders of Comput System Performance,Proc 10th Int Symp,Paris,Amsterdam:Elsevier,1984:507~519

5.Marsan M A,Donatelli S,Neri F et al.On the construction of abstract GSPNs:an exercise in modeling[C].In:the proceedings of the fourth international workshop on Petri nets and performance models,Melbourne,Australia,1991-12:2~17

6.王晖,刘卫东,杨春胜.基于PETRI网的工作流模型分析与应用[J].计算机工程与应用,2003;39(6):100~102

7.林闯,曲阳,郑波等.一种随机Petri网性能等价化简与分析方法[J].电子学报,2002;30(11)

8.田立勤,林闯,周文江.随机Petri网模型中变迁的串、并联性能等价化简技术[J].电子学报,2002;30(8)