第 19 卷 第 11 期　　控 制 与 决 策　　2004 年 11 月
Vol 19 No. 11　　Control and Decision　　Nov. 2004

文章编号: 1001-0920(2004)11-1305-03

基于核聚类方法的多层次支持向量机分类树

张国宣, 孔　锐, 施泽生, 郭　立, 刘士建, 薛明东
(中国科学技术大学 电子科学与技术系, 安徽 合肥 230026)

摘　要: 针对解决多类模式识别问题的 SVM 方法进行研究, 在比较几种常用的多类 SVM 分类算法的基础上, 提出一种基于核聚类方法的多层次 SVM 分类树, 将核空间中的无监督学习方法和有监督学习方法结合起来, 实现了一种结构更加简洁清晰、计算效率更高的多层 SVM 分类树算法, 并在实验中取得了良好的结果

关键词: 多类模式识别; 支持向量机; 核聚类; 统计学习理论
中图分类号: TP391.41　　**文献标识码**: A

Hierarchical support vector machines based on kernel clustering

ZHANG Guo-xuan, KONG Rui, SHI Ze-sheng, GUO Li, LIU Shi-jian, XUE Ming-dong
(Department of Electronic Science and Technology, University of Technology and Science, Hefei 230026, Correspondent: ZHANG Guo-xuan, E-mail: zgx@ustc.edu)

Abstract: The support vector machines (SVM) for multiclass pattern recognition are discussed. Some common SVMs for multiclass classification problems are compared. In particular, a hierarchical support vector machine is proposed based on kernel clustering method and combine the unsupervised learning method and supervised learning together. The algorithm is more effective and simple in structure and proved to performance better by experiment.

Key words: multiclass pattern recognition; support vector machine; kernel clustering; statistical learning theory

1　引　　言

由Vapnik 提出的统计学习理论系统地阐述了从样本中进行学习的方法[1,2], 并描述了有限样本情况下的统计学习问题, 首次提出了解决两类分类问题的支持向量机(SVM) 方法. SVM 方法建立在统计学习理论的VC 维理论和结构风险最小化原则基础上, 根据有限样本信息在模型的复杂性(对特定训练样本的学习精度) 和学习能力(无错误地识别任意样本的能力) 之间寻求最佳折衷, 获得最好的推广能力. 与传统的模式识别方法(如神经网络、参数估计方法等) 相比, 支持向量机在理论基础和算法性能上都表现出很大的优势. 其基本思想可概括为: 首先通过非线性变换将输入空间变换到一个高维空间, 然后在这个新空间中求取最优线性分类面, 而这种非线性变换是通过定义适当的内积函数 $K(x_i, x_j)$ 实现的.

本文提出一种新的基于核聚类方法预定义子任务的多层次 SVM 分类树. 首先介绍了多层 SVM 分类树, 然后提出一种新的多层次 SVM 分类树算法, 并通过实验对算法进行评价, 取得了良好的结果.

2　算法简介

SVM 算法是在统计学理论基础上发展起来的一种新的模式识别方法. SVM 分类器的核心目的是找到两类样本间的最优分类面. 假设样本集为 (x_i, y_i), 其中: $i = 1, \cdots, n$, $x \in R^d$, $y \in \{+1, -1\}$

收稿日期: 2003-12-31; **修回日期**: 2004-02-25.
基金项目: 教育部高校博士点基金资助项目(20020358023).
作者简介: 张国宣(1977—), 女, 安徽人, 博士生, 从事信号与信息处理、模式识别等研究; 施泽生(1937—), 男, 教授, 博士生导师, 从事图像识别与人工智能、统计学习理论等研究.

为类别符号. SVM 通过解一个不等式约束下的二次函数极值问题[2,3], 最终得到最优分类函数

$$f(x) = \mathrm{sgn}\{(w^* \cdot x) + b^*\} = \mathrm{sgn}\left\{\sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^*\right\}. \quad (1)$$

在解决实际问题时, 需要处理的大部分是多类模式识别问题, 如语音、字符识别等系统. 基本的 SVM 分类器只是两类分类器, 为将其应用到多类识别问题, 人们研究了多种解决方法. 目前 SVM 解决多类识别问题的方法一般分为标准算法和构造多层分类树方法.

标准算法包括 1-a-r 和 1-a-1 算法. 对于 N 类问题, 二者分别需要构造 N 个和 $N(N-1)/2$ 个分类器, 它们的推广误差都无界. 1-a-1 算法可能因单个分类器不规范化而趋向于过学习. Platt 提出决策导向循环图(DDAG), 将多个两类分类器组合成多类分类器, 需要 $N(N-1)/2$ 个两类分类器. 其优点是推广误差只取决于类数 N 和节点上的类间间隙, 速度比标准算法快. DDAG 和 1-a-1 的共同缺点是分类器数目随类数增加而急剧增加, 决策时速度很慢. Weaton 提出推广优化问题的 qp-mc-sv 和线性规划 lp-mc-sv 算法, 在构造决策函数时, 同时考虑所有类, 推广原始优化问题. 它们的缺点是计算量都很大, 优点是得到的决策分类面所需 SVM 的数量比常规方法少[5].

构造多层分类树, 方法是分类树中的每个节点都是完成一个预定义的分类子任务的两类分类器[4]. 这类方法分类器的结构简单清晰, 可根据样本特征向量的特性定义子任务, 推广误差只取决于类和节点上的类间间隙. 同时, 算法所需的基本两类分类器数目小于上述算法(特别是 1-a-1 和 DDAG 算法) 的数目, 而且也不需要计算庞大的拓展优化问题, 因而算法的速度较快. 可见, 只要能合理地定义分类子任务的层次, 多层分类树算法比上述算法具有明显的优势.

3 多层次 SVM 分类树

多类模式识别的重要问题之一就是模糊类. 它是指所有 1～ N 类的一个子集, 其中样本特征向量间有相似之处, 被度量的特征中有少量误差就可能导致误分类, 如在 OCR 中, 通常定义 $\{o, 0, O, Q\}$ 为一个模糊类, 因为其成员的特征向量都很相似. 多层分类器的主要思想是首先在模糊类之间进行一个粗略的分类, 然后在模糊类中进行精细的分类[4].

建立多类分类问题的多层模糊类结构通常都需要预定义, 且当一个类的集合划分为两个不相交的子类时, 必须始终通过某种误差函数的评价, 直到所有模糊类中都只包括一个独立的类. 为决定模糊类结构矩阵, 需要解整个 N 类分类问题, 同时计算所有可能的分类方法, 而可能的两类分法为 $2^{N-1}-1$, 计算量很大. 对于较大的 N, 对所有可能的两类分类的穷举更难处理, 所以模糊类结构矩阵也只能用启发式的方法计算.

多层分类器中分类子任务的定义是多层分类树设计中的核心环节, 它直接关系到分类树的效率和性能. 上文提到的定义子任务过程中的问题明显影响了算法的实用性. 为克服这些缺点, 本文设计了基于核聚类算法的多层 SVM 分类树. 下面对该算法进行描述.

4 基于核聚类的多层 SVM 分类树

不同于第 2 节所述的多层模糊类结构的分类树, 本节采用的设计多层分类树的方法是将 K 个类别首先通过核聚类的方法分为两个子类 K_1 和 K_2, 再用同样的核聚类方法将 K_1 和 K_2 分别分为两个子类, 如此将每个子类都用聚类的方法分为两个子类, 直到每个子类中只包括一个独立的类. 据此确定分类树的层次结构. Friedhelm 使用 C 均值聚类设计分类树中的子任务, 但没有取得算法总体性能上的明显改善[4].

事实上, SVM 算法的独特之处在于它将输入空间中非线性可分的样本映射到高维空间, 使其变得线性可分. 向高维空间的非线性映射, 实际上起到了分辨、提取并放大有用特征的作用. 因此, 本文从预分类的阶段便引入非线性映射, 在预定义子分类任务时采用基于 Mercer 核的聚类方法(简称核聚类方法), 应用统计学习理论的有关思想, 使整个算法的过程实际上在高维空间完成, 使其性能得到提高.

基于 Mercer 核的聚类方法是 Girolami[3] 提出的一种基于统计学习理论的无监督学习方法. 其基本思想类似于 SVM, 利用 Mercer 核函数, 首先将输入空间的样本映射到高维特征空间, 再在高维空间中进行聚类. 这种方法在性能上明显优于传统的 C 均值方法或模糊 C 均值方法. 根据统计学习理论, 由于向高维空间进行非线性映射对特征的分辨和凸显作用, 核聚类方法实现了更为准确的聚类, 且算法收敛速度更快. 在经典算法失效的情况下, 核聚类算法也能实现准确的聚类.

在核 C 均值聚类算法中, 设输入空间样本为 $x_k \in R^L$, $k = 1, \cdots, l$, 由某种非线性映射 Φ 映射到某一

高维特征空间 H 得到 $\Phi(x_k)$, 则输入空间的点积在特征空间可用Mercer核表示为 $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. 对应所有样本组成核函数矩阵 $K_{L\times L}$, $K_{i,j} = K(x_i, x_j)$. 高维特征空间中的距离可定义为

$$d_h(x_i, x_j) = (K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j))^{1/2}, \quad (2)$$

聚类准则为最小化目标函数

$$J = \sum_{i=1}^{C} \sum_{\substack{j=1 \\ x\in C_i}} \left[K(x_j, x_j) - \frac{2}{N_t} \sum_{k=1}^{N_i} K(x_j, x_k) + \frac{1}{N_{i_k,p=1}^{2}} \sum_{1}^{N_t} K(x_k, x_p) \right] \quad (3)$$

其中: C 为聚类个数, N_i 为第 C_i 类样本个数, 该类中心的模为

$$\| W_i \|^2 = \frac{1}{N_{i_k,p=1}^{2}} \sum_{1}^{N_t} K(x_k, x_p).$$

设有 N ($N > 2$) 类分类问题, 基于核聚类的多层次 SVM 分类树算法的基本步骤如下:

Step 1: 将 N 类的所有原型样本 (x_λ, y_λ), $\lambda = 1, \cdots, M$, $y_\lambda = 1, \cdots, N$, 通过两个聚类中心的核 C 均值聚类方法, 分为两个子类 K_1 和 K_2, 此时全部 N 类样本的集合为分类树的根节点, K_1 和 K_2 为第 1 层的两个叶节点

具体方法为: 设对 N 类 M 个样本的核 C 均值聚类的直接结果为两个聚类 C_1 和 C_2, 类别 i 的定义为 $\Psi_i = \{(x_\lambda, y_\lambda) \mid \lambda = 1, \cdots, M, y_\lambda = i\}$, 类 i 和聚类 j 中成员的相对频率为

$$p_{ij} := |\Psi_i \cap C_j| / |C_j|,$$

则对应于聚类 C_1 和 C_2, 子类 K_1 和 K_2 可通过下式分配:

$$K_j = \{(x_\lambda, y_j) \mid \lambda = 1, \cdots, M, y_j = \arg\max\{p_{i1}, p_{i2}\}\}, j = 1, 2 \quad (4)$$

Step 2: 对每个子类 K_i 重复应用 Step 1, 记录子类所在的层次, 直到每个子类中都只包含一个单独的类

Step 3: 根据 Step 1 和 Step 2 中对各子类预分类所得的结果及其层次, 定义分类子任务, 建立多层 SVM 分类树. 树中的每个节点对应于一个两类 SVM 分类器

这样, 算法中所有数据样本的预分类、分类子任务的预定义及分类树的建立, 都是映射到高维空间完成的, 统计学习理论保证了这一非线性映射过程, 使样本之间的特征区别更明显、分类效果更优

5　算法评价

采用MNIST 数据库对算法进行测试, 并与传统的算法进行比较. 所采用的MNIST 数据库中包含了15 000 个手写体数字(0～9) 的样本, 是一个典型的 N ($N = 10$) 类多类模式识别问题. 将所有样本分为包含 10 000 个样本的训练集和包含 5 000 个样本的测试集. 实验中采用的是径向基核函数, 构造的分类树如图 1 所示

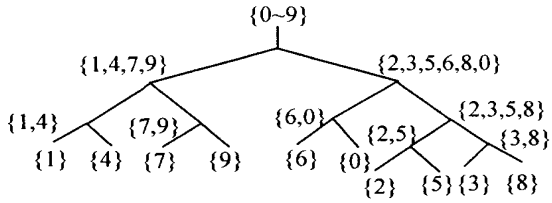

图 1　基于核聚类的 SVM 分类树(对MNIST 数据库)

表 1　实验结果比较(N = 10)

分类器	两类 SVM	错误率/%
本文算法	$N - 1$	1.34
原始多层 SVM 树	$N - 1$	1.50
Friedhelm 方法	$N - 1$	1.41
1-a-r 算法	N	1.39
1-a-1 算法	$N(N-1)/2$	1.35
DDAG	$N(N-1)/2$	1.37
lp-mc-sv	拓展优化函数	1.43

表 1 给出了本文方法与几种 SVM 分类树及常用多类 SVM 算法在基本两类 SVM 分类器个数和分类错误率的比较结果, 因为这两个指标直接反映了算法的运算效率和分类性能. 实验结果证明, 在多层分类树算法中, 本文算法相对原始多层 SVM 分类树和 Friedhelm 方法在正确率上有明显提高; 在所有的多类分类器中, 本文算法所需两类 SVM 分类器的数目最少, 但却达到了与其他算法相同的正确率. 说明本文算法实现了更为准确的分类子任务层次的定义, 使算法的结构更加简洁, 效率更高, 同时具有优良的分类性能

6　结　　语

本文针对多类模式识别问题的 SVM 方法进行研究, 提出一种基于核聚类方法的多层次 SVM 分类树. 这样可将核空间中的无监督学习和有监督学习方法结合起来, 将整个多类分类问题映射到高维空间中解决, 使得对多层次分类树子任务的定义更准确有效, 实现了一种结构更简洁、计算更有效的多层 SVM 分类树算法, 并在应用实验中取得了良好的效果, 体现出良好的性能. 进一步的工作是对各种多类 SVM 算法的深入研究

(下转第 1311 页)

进行计算, 得到算法的结果如表 2 所示

表 2　不同种群下算法 50 次运算的统计结果

种群数目	最大值	平均值	方差	一次运算时间/s
1 000	1 023.030	887.834	46.920	5
2 000	1 055.608	900.255	48.158	10
3 000	1 020.766	903.199	47.945	17
5 000	1 063.420	905.982	44.460	36

从表 2 可以看出, 种群增大可以使最好解的平均值增大, 说明种群数目增大可以使算法获得最好解的机会增大. 算法结果的方差始终较小, 说明算法的收敛比较平稳. 另外, 算法的运算时间较短.

4　结　　论

UCAV 的任务规划问题是一个大规模的复杂的优化问题. 本文建立了UCAV 任务规划问题的数学模型, 提出了分层递阶的任务规划系统结构. 针对任务规划的核心问题资源调度, 建立了数学模型, 采用遗传算法设计了动态资源调度算法, 有效地解决了UCAV 任务规划中的资源调度问题. 从计算案例结果可以看出, 算法结果科学合理, 收敛速度快且收敛平稳, 是一种快速有效的动态资源调度算法.

参考文献(References):

[1] Chandler P R, Pachter M. Research issues in autonomous control of tactical UAVs[A]. *Proc of the American Control Conf*[C]. Philadelphia, 1998. 394-398.

[2] Jovan D Boskovic, Ravi Prasanth, Raman K Mehra. A multi-layer architecture for intelligent control of unmanned aerial vehicles[A]. *AIAA*[C]. 2002. 3473.

[3] Lee Zne jung, Su Shun feng, Lee Chou yuan. Efficiently solving general weapon-target assignment problem by genetic algorithms with greedy eugenics[J]. *IEEE Trans on Systems, Man and Cybernetics*, 2003, 33(1): 113-121.

[4] Abrahams P, Balart R, Byrnes J S, et al. MAAP: The military aircraft allocation planner[A]. *The* 1998 *IEEE Int Conf on Evolutionary Computation Proc*[C]. Anchorage, 1998. 336-341.

[5] Sushil J Louis, John McDonnell, Nick Gizzi. Dynamic strike force asset allocation using genetic search and case-based reasoning[A]. *Proc of the Sixth Conf on Systems, Cybernetics, and Informatics*[C]. Orlando, 2002. 855-861.

[6] Balart R, Byrnes J S, Cochran D, et al. Decision aids for asset-to-objective allocation[A]. 1995 *Conf Record of the Twenty-Ninth Asilamar Conf on Signals, Systems and Computers*[C]. Pacific Grore, 1995. 807-811.

[7] 黄俊, 孙义东, 武哲, 等. 战斗机对地攻击作战效能分析[J]. 北京航空航天大学学报, 2002, 28(3): 354-357.
(Huang J, Sun Y D, Wu Z, et al. Operational effectiveness analyses of air-to-ground strike for battle-plan[J]. *J of Beijing University of Aeronautics and Astronautics*, 2002, 28(3): 354-357.)

[8] 邵力军, 张景, 魏长华. 人工智能基础[M]. 北京: 电子工业出版社. 2001. 203-219.

(上接第 1307 页)

参考文献(References):

[1] 边肇祺. 模式识别[M]. 北京: 清华大学出版社, 1999. 294-304.

[2] Vapnik. 统计学习理论的本质[M]. 张学工译. 北京: 清华大学出版社, 2000.

[3] Mark Girolami. Mercer kernel based clustering in feature space[J]. *IEEE Trans on Neural Networks*, 2002, 5(13): 780-784.

[4] Friedhelm Schwenker. Hierarchical support vector machines for multi-class pattern recognition[A]. *Fourth Int Conf on Knowledge-based Intelligent Engineering Systems & Allied Technologies*[C]. Brighton, 2000. 561-565.

[5] Platt J C, Cristianini N, Shawe Taylor J. Large margin DAGs for multiclass classification[A]. *Advances in Neural Information Processing Systems*[C]. MIT Press, 2000. 547-553.