文章编号: 1001-0920(2005)08-0931-04

一种新的分裂层次聚类 SVM 多值分类器

张国云[1,2], 章 兢[1]
(1. 湖南大学 电气与信息工程学院, 长沙 410082; 2. 湖南理工学院 电子信息系, 湖南 岳阳 414006)

摘 要: 提出一种分裂层次聚类 SVM 分类树分类方法. 该方法通过融合模糊聚类技术和支持向量机算法, 利用分裂的层次聚类策略, 有选择地重新构造学习样本集和 SVM 子分类器, 得到了一种树形多值分类器. 研究结果表明, 对于 k 类别模式识别问题, 该方法只需构造 $k-1$ 个 SVM 子分类器, 克服了 SVM 子分类器过多以及存在不可区分区域的缺点, 具有良好的分类性能. 实验结果验证了该方法的优越性.

关键词: 分裂层次聚类; 支持向量机; 多值分类器; 分类树
中图分类号: TP18 **文献标识码**: A

A Novel SVM Multi-class Classifier Based on Divisive Hierarchical Clustering

ZHANG Guo-yun[1,2], *ZHANG Jing*[1]
(1. College of Electrical and Information Engineering, Hu'nan University, Changsha 410082, China; 2. Department of Electronics Information, Hu'nan Institute of Science and Technology, Yueyang 414006, China. Correspondent: ZHANG Guo-yun, E-mail: gyzhang2283@sina.com)

Abstract: A novel support vector machines classification tree approach based on divisive hierarchical clustering (DHCSVC) is proposed. By fusion of fuzzy clustering techniques and support vector machines (SVM), and utilizing the divisive hierarchical clustering strategy, this method selectively re-constructs learning samples set and corresponding SVM sub-classifier. A tree-shape multi-class classifier is also built using this method. In k-class task, the new classifier only contains $k-1$ SVM sub-classifiers. The proposed method also overcomes the drawbacks such as unclassifiable region which the "one-against-one" method has, and possesses a good classification performance. The experimental results show that the proposed approach has superiority in classification task.

Key words: Divisive hierarchical clustering; SVM; Multi-class classifier; Classification tree

1 引 言

支持向量机(SVM)[1,2]是基于统计学习理论的一种新的机器学习方法, 具有严格的数学基础. 由于它是建立在结构风险最小化准则上, 使得支持向量机分类器具有较好的推广能力, 在模式识别中得到了实际应用[3]. 然而, 支持向量机是针对二类别问题提出的, 如何将二类别分类扩展到多类别分类, 是支持向量机研究的重要内容之一.

构造 SVM 多类别分类方法主要有两类[4]: 一类是在经典 SVM 理论的基础上, 重新构造多类别分类模型, 通过 SV 方法对新分类模型目标函数进行优化, 实现多类别分类, 但算法所选择的目标函数复杂, 实现困难[5,6]; 另一类的基本思想是通过组合多个二类别子分类器, 实现多类别分类器的构造. 常用的构造方法有"一对一"和"一对多"两种, 一对一方法分类精度高于一对多方法[7]. 然而, 一对一方法存在不可区分区域[8], 对于 k 类别问题, 需要 $k(k-1)/2$ 个子分类器, 子分类器数目较大, 需要内存较多.

国内研究人员近几年提出了几种新的 SVM 多

收稿日期: 2004-09-22; **修回日期**: 2004-12-08.
基金项目: 教育部科学技术研究重点项目([2001]224).
作者简介: 张国云(1971—), 男, 湖南桂阳人, 副教授, 博士生, 从事模式识别、智能控制等研究; 章兢(1957—), 男, 湖南湘潭人, 教授, 博士生导师, 从事模式识别、人工智能等研究.

值分类器[9, 10], 均表现在把 SVM 与其他方法相结合. 这些方法在一定程度上提高了分类精度和分类速度, 但构造多值分类器时均用到了一对一方法, 即需构造 $k(k-1)/2$ 个子分类器, 也需较多的计算时间和内存开销. 针对这一问题, 本文通过融合模糊聚类技术[11] 和支持向量机算法, 利用分裂的层次聚类策略[12], 提出一种树形多值分类器.

2 支持向量机

支持向量机(SVM) 通过寻求结构风险最小, 根据有限样本信息, 在模型的复杂性与学习能力之间寻求最佳折衷, 实现经验风险和置信范围的最小化, 从而在统计样本量较少的情况下, 获得更好的泛化能力和良好的统计规律. 根据泛函中的Mercer 定理, SVM 采用核函数, 在高维线性特征空间中构造最优分类超平面, 得到分类器的决策函数.

假设存在训练样本 $(x_1, y_1), \cdots, (x_l, y_l)$, $x \in R^n$, $y \in \{+1, -1\}$. 其中 l 为样本数, n 为输入维数. 当训练样本集线性不可分时, 引入非负松弛变量 ξ_i, $i = 1, 2, \cdots, l$. 分类超平面最优化问题描述为

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i;$$

$$\text{s.t. } y_i(w^{\mathrm{T}}\phi(x_i) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0,\ i = 1, 2, \cdots, l. \tag{1}$$

通过求解最优化问题, 可得到相应的最优决策函数

$$f(x) = \operatorname{sgn}\Big(\sum_{i=1}^{l} y_i\alpha_i k(x, x_i) + b\Big). \tag{2}$$

3 分裂层次聚类 SVM 多值分类器

对于 k 类别分类问题, 为提高分类效果, 本文通过集成模糊聚类技术与 SVM 二值分类器, 利用分裂的层次聚类策略, 提出一种新的分裂层次聚类 SVM 多值分类器(DHCSVC), 即分裂层次聚类 SVM 分类树.

3.1 分裂层次聚类 SVM 分类树

定义 1 满足下列条件的 5 元组 DHCSVC = $\langle U, S, F, \mathrm{New}, \mathrm{SVM}\rangle$ 称为分裂层次聚类 SVM 分类树.

1) $F = \bigcup_{i=1}^{k} F_i$, $F = \{(x_1, y_1), \cdots, (x_l, y_l)\}$, $F_i = \{(x_1, y_i), \cdots, (x_{m_i}, y_i)\}$, $x_j \in R^n$, $y_i \in \{1, 2, \cdots, k\}$, $\sum_{i=1}^{k} m_i = l$. 其中 F 为学习样本集, F_i 为 F 的 k 个划分.

2) $U = \{U_1, U_2, \cdots, U_k\}$, U_i 为 F_i 的聚类中心点, $U_i \in R^n$.

3) 二元组 (U, S) 组成二叉树, S 是 U 上某个二元关系 H 的集合, (U_{l_i}, H_{l_i}) 为根 r_i 的左子树, (U_{R_i}, H_{R_i}) 为根 r_i 的右子树, $H_{l_i} \subset H$, $H_{R_i} \subset H$.

4) $\mathrm{New} = \{\mathrm{New}_1, \cdots, \mathrm{New}_{k-1}\}$, New_i 为第 i 个子分类器 SVM_{ci} 学习样本.

5) $\mathrm{SVM} = \{\mathrm{SVM}_{c1}, \cdots, \mathrm{SVM}_{c,k-1}\}$, SVM_{ci} 为第 i 个结点处 SVM 子分类器.

根据上述定义, DHCSVC 的学习过程实际上是构造 $k-1$ 个 SVM 子分类器. 首先, 利用模糊聚类技术求取每类样本的聚类中心 $U = \{U_1, U_2, \cdots, U_k\}$, 得到根结点, 根据分裂的层次聚类策略将它们聚类成两类, 并将各聚类中心对应的样本数据分别记为正类 P_1(即左子树) 和负类 N_1(即右子树), 构造新的学习样本集 New_1 及第 1 个 SVM 子分类器 SVM_{c1}. 然后, 将 P_1 对应的聚类中心点又聚类成两类, 并将各聚类中心对应的样本数据分别记为正类 P_2 和负类 N_2, 构造第 2 个子分类器, 对负类 N_1 采取相同作法; 再对 P_2 和 N_2 对应的各聚类中心分别聚类成两类, 用同样的办法构造第 3 级子分类器. 依次下去, 直到每个子类仅含一个聚类中心点. 这些子分类器形成了一个树状结构, 在每个结点处构造一个 SVM 子分类器.

定理 1 对于 k 类别模式分类问题, DHCSVC 包含 $k-1$ 个 SVM 子分类器.

证明 对于 k 类别问题(即 DHCSVC 终端结点数为 k), 二叉树中所有结点的度均小于或等于 2. 设度为 1 的结点数为 n_1, 度为 2 的结点数为 n_2, 结点总数 $n = k + n_1 + n_2$. 在分类树中, 除根结点外, 其余结点都有一个分支进入. 设分支总数为 B, 则 $n = B + 1$. 这些分支是由度为 1 或 2 的结点射出的, 故 $B = n_1 + 2n_2$, 即 $n = n_1 + 2n_2 + 1$, 由此 $n_2 = k - 1$. 每个度为 2 的结点代表一个 SVM 子分类器, 所以 DHCSVC 包含 $k-1$ 个 SVM 子分类器. □

在 DHCSVC 构造的 $k-1$ 个 SVM 子分类器中, 下一级子分类器比上一级子分类器学习样本少, 因而所需计算时间呈递减态势. DHCSVC 根据每类样本聚类中心把相应样本数据分成两类, 这在很大程度上避免了个别边界样本点错分的可能.

3.2 SVM 子分类器的构造

构造第 i 个 SVM 子分类器 SVM_{ci} 时, 将左子树对应的样本定义为正类 P_i, 将右子树对应的样本定义为负类 N_i, $\mathrm{New}_i = P_i \cup N_i$. 不失一般性, 假设样本为线性不可分情形, 则第 i 个 SVM_{ci} 对应的最优化问题为

$$\min \frac{1}{2}\|w_i\|^2 + C\sum_{j=1}^{l_i}\xi_j^i;$$

$$\text{s.t. } w_i^{\mathrm{T}}\phi(x_j) + b_i \geq 1 - \xi_j^i,\ y_i \in P_i,$$

$$w_i^{\mathrm{T}}\phi(x_j) + b_i \leq \xi_j^i - 1, y_i \in N_i,$$
$$\xi_j^i \geq 0, i = 1, \cdots, k-1. \quad (3)$$

式中: l_i 为 New$_i$ 中样本个数, $\phi(\cdot)$ 为输入空间到高维特征空间的非线性映射函数 DHCSVC 需求解 $k-1$ 个类似式(3) 的最优化问题, 因而有 $k-1$ 个决策函数 $f_1(x), \cdots, f_{k-1}(x), f_i(x)$ 由式(2) 得到

3.3 DHCSVC 算法

下面给出DHCSVC 学习算法描述:

Step 1: 求取每类学习样本模糊聚类中心 $U = \{U_1, U_2, \cdots, U_k\}$, 设有 k 类, 则每类对应一个聚类中心.

Step 2: 利用模糊聚类技术将 U 聚类成两类 U_P 和 U_N, $U_P \subset U$, $U_N \subset U$, $U_P \cap U_N = \varnothing$, $U_P \cup U_N = U$;

Step 3: 将属于 U_P 的各聚类中心对应的所有学习样本设置为正类 P$_1$, 将属于 U_N 的各聚类中心对应的所有学习样本设置为负类 N$_1$, 重新组合学习样本得到 New$_1$, P$_1 \cup$ N$_1$ = New$_1$, P$_1 \cup$ N$_1$ = F, 构造 SVM 子分类器 SVM$_{c1}$.

Step 4: 将 U_P 聚类成两类, 其各聚类中心对应的样本记为正类 P$_2$ 和负类 N$_2$; 将 U_N 聚类成两类, 其各聚类中心对应的样本记为正类 P$_3$ 和负类 N$_3$. 其中 P$_2 \cup$ N$_2$ = P$_1$, P$_2 \cap$ N$_2$ = $\varnothing$, P$_2 \subset$ P$_1$, N$_2 \subset$ P$_1$; P$_3 \cup$ N$_3$ = N$_1$, P$_3 \cap$ N$_3$ = $\varnothing$, P$_3 \subset$ N$_1$, N$_3 \subset$ N$_1$.

Step 5: 根据 P$_2$ 和 N$_2$ 构造 SVM 子分类器 SVM$_{c2}$, 根据 P$_3$ 和 N$_3$ 构造子分类器 SVM$_{c3}$.

Step 6: 重复 Step 4 和 Step 5, 直至构造出第 $k-1$ 个 SVM 子分类器 SVM$_{c,k-1}$.

Step 7: 由 Step 1～ Step 6 得到 $k-1$ 个 SVM 子分类器 SVM$_{c1}$, ⋯, SVM$_{c,k-1}$, 相应的 $k-1$ 个新的学习样本集 New$_1$, ⋯, New$_{k-1}$, 以及 $k-1$ 个决策函数 $f_1(x), \cdots, f_{k-1}(x)$.

对测试样本测试时, 首先从根结点 SVM$_{c1}$ 子分类器开始, 判别其输出是正类还是负类, 分别表示为 +1 和 -1; 再根据输出结果, 利用相应的第 2 级 SVM 子分类器进行测试. 依次计算下去, 直至最后一级子分类器, 从而得出测试数据所属的类别.

4 实验结果

作者利用 Matlab 6.5 自行编写了整套 SVM 软件, 并在 PⅢM 1.2 G/256 M/30 G 硬件环境下运行. 实验中的模糊聚类技术直接使用 FCM 函数, 并选择加权参数为 2, 迭代步数为 200. 为测试 DHCSVC 的分类效果, 选择 Image Segmentation Dataset 数据集. 该数据集包含 210 个学习样本和 2 100 个测试样本, 分为 7 类, 每个样本含 19 个属性.

利用本文提出的DHCSVC 方法和 FCM 模糊聚类技术, 得到 7 个聚类中心点 $U_1, U_2, \cdots, U_7$, 所得的分裂层次聚类树形结构如图 1 所示. 图中 T_i 的下标代表子分类器序号, $T_1 = \{U_1, U_2, \cdots, U_7\}$, $T_2 = \{U_1, U_3, U_4, U_5, U_6, U_7\}$, $T_3 = \{U_4, U_6\}$, $T_4 = \{U_1, U_3, U_5, U_7\}$, $T_5 = \{U_1, U_3, U_5\}$, $T_6 = \{U_1, U_5\}$.

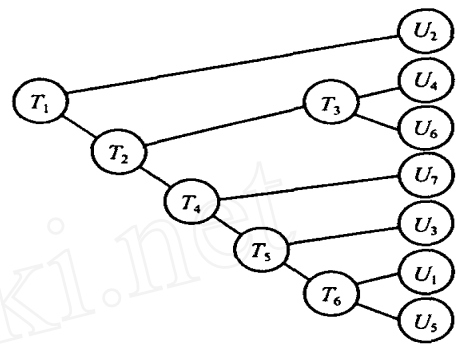

图 1　聚类中心点分裂层次聚类树形结构

重新构造 6 个学习样本集, 即 New = {New$_1$, ⋯, New$_6$}, 以及 6 个 SVM 子分类器 SVM$_{c1}$, ⋯, SVM$_{c6}$. 子分类器选取 RBF 核函数, 惩罚参数 C = 100, 停止误差 ϵ = 0.001. 图 2 给出了DHCSVC 方法和一对一方法在 σ 取不同值时, 对 2 100 个测试样本的分类正确率.

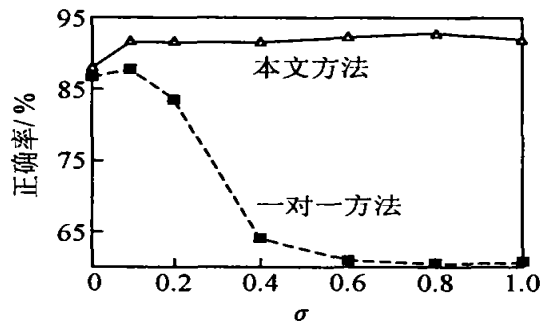

图 2　σ 取不同值的分类正确率

选取 σ = 0.1, 惩罚参数 C = 100, 停止误差 ϵ = 0.001, 对不同规模测试样本集的分类正确率如图 3 所示.

当 σ = 0.1, 惩罚参数 C = 100, 停止误差 ϵ =

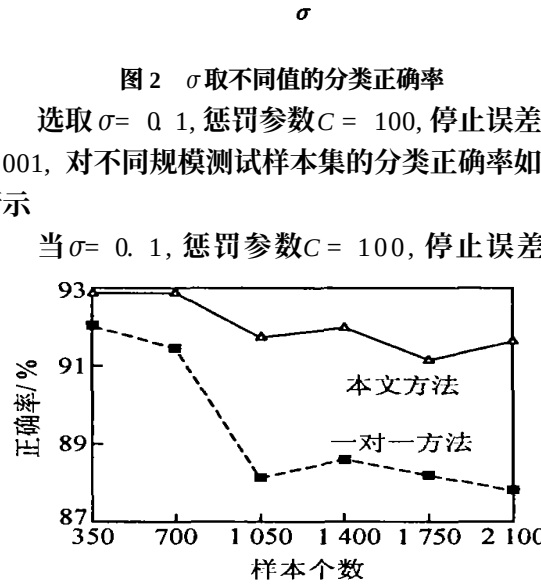

图 3　不同规模测试样本集分类正确率

0.001 时, 对不同规模样本集测试所需分类时间如图 4 所示, DHCSVC 所需分类时间明显要少. 这是因为一对一方法需 21 个子分类器, 而 DHCSVC 只需 6 个子分类器. 实验中根据不同规模的测试样本集, 对每类测试样本按顺序分别选取 50, 100, 150, 200, 250, 300 个样本, 从而可得到样本规模分别为 350～2 100 的 7 种情形.

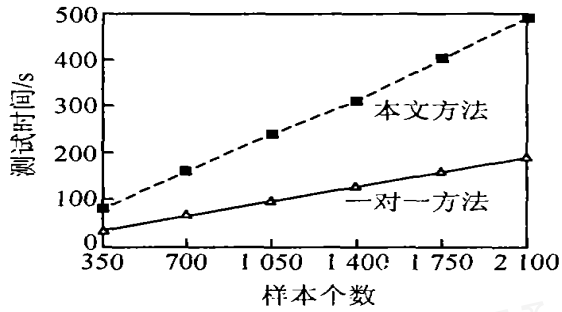

图 4 不同规模测试样本集测试时间

由实验结果可知, DHCSVC 分类精度优于一对一方法. 主要原因是 DHCSVC 在学习过程中集成了模糊聚类技术和 SVM 算法, 它包含粗分和细分两个过程, 即先用模糊聚类技术对样本粗分, 有选择地构造 SVM 子分类器, 再用 SVM 子分类器对样本实现精细分类.

5 结 论

本文提出一种分裂层次聚类 SVM 多值分类器(DHCSVC). 该方法通过求取每类模糊聚类中心并对聚类中心进行分裂, 在层次聚类的基础上, 有选择地重新构造 $k-1$ 个学习样本集, 进而构造 $k-1$ 个 SVM 子分类器, 在很大程度上避免了边界样本点错分的可能. 理论证明 DHCSVC 仅需 $k-1$ 个子分类器, 因而只需构造较少的子分类器. 实验结果表明, DHCSVC 分类精度优于一对一方法, 测试过程速度加快. 由于采用了分类树结构, 不存在不可区分区域问题.

参考文献(References)

[1] 张学工. 关于统计学习理论与支持向量机[J]. *自动化学报*, 2000, 26(1): 32-42.
(Zhang X G. Introduction to Statistical Learning Theory and Support Vector Machines[J]. *Acta Automatica Sinica*, 2000, 26(1): 32-42.)

[2] Vapnik V. *The Nature of Statistical Learning Theory* [M]. New York: Springer-Verlag, 1995.

[3] Burges C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. *Data Mining and Knowledge Discovery*, 1998, 2(2): 121-127.

[4] 萧嵘, 孙晨, 王继成, 等. 一种具有容噪性能的 SVM 多值分类器[J]. *计算机研究与发展*, 2000, 37(9): 1071-1075.
(Xiao R, Sun C, Wang J C, et al. A Noise Insensitive SVM Multi-class Classifier[J]. *J of Computer Research and Development*, 2000, 37(9): 1071-1075.)

[5] Xin D, Wu Z H, Pan Y H. A New Multi-class Support Vector Machines[A]. *Proc of IEEE Int Conf on System, Man and Cybernetics*[C]. New York: 2001, 3: 1673-1676.

[6] Arenas Garcia J, Perez Cruz F. Multi-class Support Vector Machines: A New Approach[A]. *Proc of the IEEE Int Conf on Acoustics, Speech and Signal Processing*[C]. New York, 2003, 2: 781-784.

[7] Hsu C W, Lin C J. A Comparison of Methods for Multi-class Support Vector Machines[J]. *IEEE Trans on Neural Networks*, 2002, 13(2): 415-425.

[8] Inoue T, Abe S. Fuzzy Support Vector Machines for Pattern Classification[A]. *Proc of Int Joint of Conf on Neural Networks*[C]. New York, 2001, 2: 1449-1454.

[9] 李蓉, 叶世伟, 史忠植. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. *电子学报*, 2002, 30(5): 745-748.
(Li R, Ye S W, Shi Z Z. SVM-KNN Classifier - A New Method of Improving the Accuracy of SVM Classifier[J]. *Acta Electronica Sinica*, 2002, 30(5): 745-748.)

[10] 李红莲, 王春花, 袁保宗. 一种改进的支持向量机 NN-SVM[J]. *计算机学报*, 2003, 26(8): 1015-1020.
(Li H L, Wang C H, Yuan B Z. An Improved SVM: NN-SVM[J]. *Chinese J of Computers*, 2003, 26(8): 1015-1020.)

[11] Pal N R, Bezdek J C. On Cluster Validity for the Fuzzy *c*-means Model[J]. *IEEE Trans on Fuzzy Systems*, 1995, 3(3): 370-379.

[12] Jiawei Han, Micheline Kamber. *Data Mining: Concepts and Techniques*[M]. San Francisco: Morgan Kaufmann Publishers, 2000.