文章编号: 1001-0920(2011)01-0149-04

一种新的基于平衡决策树的SVM多类分类算法

刁智华[1,2], 赵春江[1], 郭新宇[1], 陆声链[1]

(1. 国家农业信息化工程技术研究中心，北京 100097；2. 中国科学技术大学 自动化系，合肥 230026)

摘 要: 为了有效地减少样本训练时间, 提高多类分类器的识别率, 同时使模型具有较好的推广能力, 在综合考虑待分类样本数和类别易分性能的基础上, 在“先分样本数较大的类”和“先分易分的类”之间折衷考虑, 提出一种基于样本的新的类划分方案. 采用平衡决策树结构, 得到了一种新的决策树支持向量机多类分类算法. 实验结果表明, 该算法在不降低识别率的情况下, 能大大减少系统的训练时间, 是一种有效的多类分类算法.

关键词: 支持向量机；决策树；多类分类器；类间可分性

中图分类号: TP394 **文献标识码:** A

A new SVM multi-class classification algorithm based on balance decision tree

DIAO Zhi-hua[1,2], *ZHAO Chun-jiang*[1], *GUO Xin-yu*[1], *LU Sheng-lian*[1]

(1. National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China; 2. Department of Automation, University of Science and Technology of China, Hefei 230026, China. Correspondent: DIAO Zhi-hua, E-mail: dzhua@mail.ustc.edu.cn)

Abstract: In order to decrease the sample training time effectively, improve the identification rate, and make the model has good generalization ability, a new class partition project based on samples is proposed. This project makes a comprehensive consideration of the number of waiting classification samples and the capability of class partition, and takes a compromise between the “first classifying the classes with a large number of samples” and the “first classifying the classes that can be partitioned easily”. And a new decision-tree-based support vector machines multi-class classification algorithm is proposed, which adopts the balance decision tree structure. The experimental results show that the algorithm can significantly reduce system training time at the condition of not reducing identification rate, and is an effective multi-class classification algorithm.

Key words: support vector machines; decision tree; multi-class classifiers; inter-class separability

1 引 言

支持向量机(SVM)是基于统计学习理论提出的解决小样本模式识别问题的有效方法[1], 其基本思想是构造一个超平面作为决策平面, 使正负类别之间的间隔最大. SVM能较好地解决非线性、高维数、局部极小点等问题, 对于特征相关性和稀疏性不敏感, 已成为继神经网络之后机器学习的一个新的研究热点.

SVM本身是一个两类分类算法, 在解决多类分类问题时会遇到难以逾越的瓶颈. 因此, 如何将其推广到多类分类问题以适应实际应用的需要是目前的研究热点. 不仅如此, 如何提高识别率、获得较好的推广能力、减少训练时间和测试时间也是值得考虑的问题. 目前, 虽然对于多类SVM的理论研究还有待于完善, 但也取得了一定成果, 已经有多种方法将SVM推广到多类分类问题, 其中主要分为整体法和分解法两大类[2]: 整体法是在所有训练样本上一次性求解一个大的二次规划问题, 同时将多类分开[3], 该方法思路虽然简单, 但在求解最优化问题的过程中使用的变量太多, 计算复杂度过高且不适用; 分解法包括用多个两类分类器来实现多类分类(如one-against-one方法[4], one-against-rest方法[5], 有向无环图方法[6])和用层次型两类分类器实现多类分类(如基于决策树的SVM方法[7-8])这两种方法. 对于多类分类问题而言, 决策树SVM算法需要构造的分类器最少, 而且具有不存在不可分区域的优点, 在分类时也不需要遍历所

收稿日期: 2009-10-15；**修回日期:** 2010-01-06.

基金项目: 国家863计划项目(2007AA10Z237)；北京市自然科学基金项目(40810010).

作者简介: 刁智华(1982–), 男, 博士, 从事图像处理、多类分类算法的研究；赵春江(1964–), 男, 研究员, 博士, 从事农业信息技术等研究.

有的分类器. 大量的研究和实验[8-9]均表明, 基于决策树的SVM多类分类算法是目前SVM解决多类分类问题中最好的方法.

为了提高多类分类器的识别率, 减少训练时间, 使模型具有较好的推广能力, 在综合考虑待分类样本数以及类别易分性能的基础上, 在“先分样本数较大的类”和“先分易分的类”之间折衷考虑, 提出了一种基于样本的新的类划分方案, 并采用平衡决策树结构, 得到一种新的SVM多类分类算法. 实验结果表明, 该算法在不降低识别率的情况下, 能大大减少系统的训练时间, 是一种有效的多类分类算法.

2 基于决策树的SVM多类分类算法

决策树算法将所有类别分为两个子类, 每个子类又划分为两个子子类, 如此循环, 直到划分出最终类别. 每次划分后两类分类问题的规模逐级下降, 这样得到一个倒立的决策树, 每个决策点用SVM实现分类. 对于m类问题, 需要构造$m-1$个分类器. 决策树分类的优点是不存在不可分区域, 分类时不需要遍历所有的分类器. 同时, 由于采用树形结构, 分类时由上向下的层次使得分类器中的支持向量数逐层减少, 训练所需的样本数也会变小, 不仅能减少训练时间, 还能较好地提高分类效率. 其缺点主要是决策树分类结构存在误差累积问题, 同时决策树的结构对其推广能力影响很大. 因此, 如何确定一个较好的决策树结构是当前学者们研究的方向. 目前, 对于决策树结构的研究主要包括非平衡决策树和平衡决策树两种.

2.1 基于非平衡决策树结构的多类分类算法

非平衡决策树的结构如图1所示, 以6类分类为例, 需要构造5个分类器, 叶子节点为样本的最终类别. 非平衡决策树的分类方法算法如下: 第1个SVM以第1类样本为正的训练样本, 将第$2,3,\cdots,m$类训练样本作为负的训练样本训练SVM; 第i个SVM以第i类样本为正的训练样本, 将第$i+1,i+2,\cdots,m$类样本作为负的训练样本训练SVM(i); 直到第$m-1$个SVM将以第$m-1$类样本作为正样本, 以第m类样本为负样本训练SVM($m-1$). 对于m类问题, 只需要$m-1$个分类器.

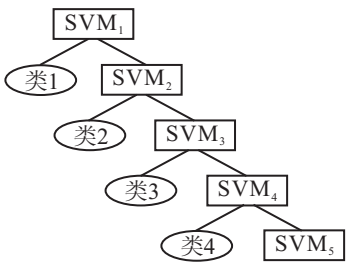

图 1 非平衡决策树结构

目前, 许多学者在利用非平衡决策树进行分类时, 非平衡决策树的结构是随机产生的. 对于非平衡决策树而言, 在树的结构已经确定的前提下, 如何根据训练样本生成分类精度更高、推广能力更好的决策树算法是研究的热点. 文献[10]在前人研究的基础上提出了基于聚类的类距离法, 在对多类数据进行聚类分析的基础上, 将与其他类距离最远或最容易分割的那一类首先分割出来, 据此标准完成对所有类的分割, 并按照类别分离出来的先后顺序, 生成一棵决策树. 但该算法无法克服野点对分类器的构造影响, 也没有考虑类内的聚合程度. [8]提出以超长方体或超球体的体积作为衡量分布区域大小的标准构造决策树, 这种方法没有考虑类间距离以及某一类数据的密度. 针对以上两种算法的改进, [11]提出了利用球结构SVM中构造超球体的方法, 对输入空间中的m类样本点构造m个球心和半径平方已知的超球体, 并将这m个超球体的空间分布作为构造决策树的依据. 这种算法有效避免了野点带来的影响, 具有一定的抗干扰性能, 但训练时间较长.

2.2 基于平衡决策树结构的多类分类算法

平衡决策树的结构如图2所示. 以6类分类为例, 需要构造5个分类器, 叶子节点为样本的最终类别. 平衡决策树的分类算法如下: 对于m类训练样本, 第1个SVM以$m/2$的整数m_1类样本为正的训练样本, 其余类训练样本作为负的训练样本训练SVM; 然后对其中1个含有m_1类样本的分支进行训练, 训练时SVM以$m_1/2$的整数m_2类样本为正的训练样本, 其余m_1-m_2类训练样本作为负的训练样本训练SVM; 对另一个分支也采用相似的方法进行训练, 直到所有类别被分类出来. 对于m类问题, 只需要$m-1$个分类器.

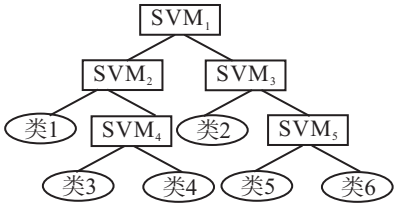

图 2 平衡决策树结构

一些学者在利用平衡决策树进行分类时, 重点研究了平衡决策树在分类器数目、分类速度方面的性能. 从树形结构中可以直观地看出, 在平衡决策树中识别出一类样本所需要的层数比非平衡决策树需要的层数少, 这样能大大减少识别时间. 当总训练样本数为m, 每个类别包含m/N个样本时, 由文献[12]得到, 非平衡决策树和平衡决策树的训练时间分别为

$$T_{Pt}=C\sum_{i=1}^{N-1}\left(\frac{N-i+1}{N}m\right)^r,$$

$$T_{Ft} = C\sum_{i=1}^{e}\left(2^{i-1}\left[\frac{m}{2^{i-1}}\right]\right)^r,\ T_{Ft} < T_{Pt}.$$

即平衡决策树的训练时间小于非平衡决策树的训练时间, 其他决策树的训练时间介于两者之间. 对于使用分类器的平均数目, 平衡决策树为 $\log_2^N$, 非平衡决策树为 $(N+1)/2 - 1/N$. 其他决策树分类器平均值介于两者之间, 平衡决策树分类时所需的分类器数目最少, 分类速度较快. 文献[13]基于球结构设计了一种平衡决策树生成算法用于多类分类问题, 利用球体的中心和半径来建立相似性度量函数, 可以更好地去除噪声或孤立点数据. 分类时也不需要遍历所有的决策节点, 其速度相当于折半查找, 使得分类器的训练和分类速度有很大的提高. 但是由于没有考虑样本数目等因素, 训练阶段的训练速度较差.

3　新型算法

为了构造一个SVM决策树分类器, 需要确定决策树的结构以及每个结点分类器的类划分方案. 即采用不同的类组合作为结点分类器的正例类集合和反例类集合, 用正例类集合和反例类集合的训练样本对结点分类器进行训练, 从而实现对这两个类集合的划分. 根据目前的研究进展, 综合考虑训练速度、分类精度和分类速度, 本文采用平衡决策树结构, 结合训练样本中样本的分布属性, 提出了一种SVM多类分类算法. 具体算法实现包括类间可分性度量和类划分方案两个方面.

3.1　类间可分性度量

类间可分性度量采用文献[14]中介绍的方法, 并结合样本属性进行计算. 具体算法实现为: 首先统计多类样本中各个类别的样本数 n_i, 其中 i 表示样本类别; 根据文献[14]中的方法计算多类分类中各类的类间分离性测度 δ_{ij}. 最终的类间可分性度量公式为

$$D_i = n_i \sum_{j=1,\ j\neq i}^{l} \delta_{ij}, \tag{1}$$

其中 l 为多分类的类别数目.

3.2　基于样本的类划分方案

类划分方案不仅要考虑当前结点分类器及其子结点分类器的类集合的可分性[15], 还要考虑各子类的样本数目, 较早地将易分的类和样本数目较多的类分开. 这样能够保证分类准确率, 还能有效提高训练速度. 根据该思想, 得到类划分方案如下:

Step 1: 根据类间可分性度量公式计算 $D_i(i = 1, 2, \cdots, l)$, 并按照从大到小的顺序进行排列, 若存在两类的度量相同, 则将样本较多的类别排在前面.

Step 2: 将序列中前1/2的整数部分的类别标记为正类, 其余类别标记为负类. 若在分类过程中, 需要将一类分开, 且子类中存在两个类别的类间可分性度量相同, 则将样本较多的类别先分开. 即将该样本较多的类别标记为正类, 其余类别标记为负类.

Step 3: 利用SVM算法进行训练, 构造一个二值分类器.

Step 4: 分类的正负样本组成的集合为决策树的两个子节点, 若子节点的样本集合中含有两个以上的类别, 则转至Step 2, 对各子节点样本类别的一半取整数个类别标记为正类, 同时该子节点样本中的其他类别标记为负类, 执行Step 3. 若个类别被分开, 则下一步进行SVM训练时删除该类别的所有样本.

Step 5: 继续执行, 直到所有类别被分开, 得到结构模型如图3所示(以7类分类为例).

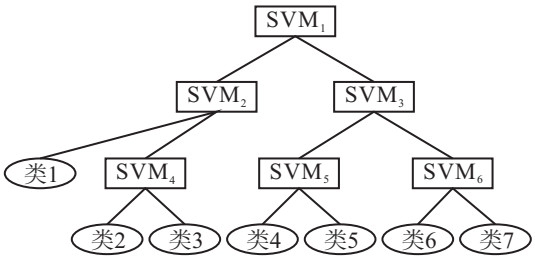

图 3　基于决策树SVM的模型结构

4　实验结果与讨论

为了比较各SVM算法的性能优势, 以Statlog[2]数据库中的多类别数据集Letter, Satimage和Shuttle作为实验对象, 分别对one-against-one, one-against-rest, 欧氏距离极大极小法[2]和新型算法进行仿真研究和比较. 表1列出了上述各数据集的相关信息. 所有算法均采用C++语言编程, 并在VC 6.0上实现, 二值SVM分类算法在SVM-Light工具包的基础上修改完成. 系统运行环境为主频2.0 GHz, 内存2 G, 操作系统为Windows XP. 在数据集训练时采用Shrinking算法来实现训练加速, 同时选用的核函数为径向基核函数, 即

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right).$$

表 1　实验数据集

数据集	训练样本集	测试样本集	类别数	属性数
Letter	15 000	5 000	26	16
Satimage	4 435	2 000	6	36
Shuttle	43 500	14 500	7	9

SVM模型采用的是 C-SVM, C 为惩罚参数, 在模型训练中需要选定参数 C, 一般采用 K-交叉验证法, 本实验只需确定核参数 σ^2 和 C . 同时, 为了避免取值范围大的属性比取值范围小的属性更占优势, 对样本数据各属性进行归一化预处理, 线性调整到 $[-1, 1]$. 经过交叉验证后, 得到上述4种算法的相对最优参

数值如表2所示. 4种算法在各自的相对最优参数值处对应的训练时间、测试时间和识别率如表3所示. 其中: $T_{\rm tr}$ 为训练时间(s), $T_{\rm test}$ 为测试时间(s), R 为识别率(%).

表 2 实验中得到的相对最优参数值

数据集	one-against-one (C,σ^2)	one-against-rest (C,σ^2)	欧式距离极大极小法 (C,σ^2)	新型算法 (C,σ^2)
Letter	$(2^7,2^{-2})$	$(2^3,2^{-2})$	$(2^3,2^{-2})$	$(2^3,2^{-2})$
Satimage	$(2^3,2^{-2})$	$(2^3,2^{-1})$	$(2^3,2^{-1})$	$(2^3,2^{-1})$
Shuttle	$(2^{12},2^{-4})$	$(2^{12},2^{-3})$	$(2^{12},2^{-4})$	$(2^{12},2^{-4})$

表 3 相对最优参数下的实验结果对比

数据集	one-against-one($T_{\rm tr}$, $T_{\rm test}$,R)	one-against-rest ($T_{\rm tr}$, $T_{\rm test}$,R)	欧式距离极大极小法 ($T_{\rm tr}$, $T_{\rm test}$,R)	新型算法 ($T_{\rm tr}$, $T_{\rm test}$,R)
Letter	(365,319,96.6)	(483,271,95.3)	(596,43,95.8)	(308,41,96.5)
Satimage	(40,29,91.3)	(53,21,90.6)	(9,8,91.7)	(8,8,91.6)
Shuttle	(458,25,97.6)	(475,21,97.4)	(537,18,98.3)	(367,18,98.4)

从表3的实验结果可以看出, 在训练时间方面, Letter数据集采用4种不同算法需要的测试时间分别为365 s, 483 s, 596 s, 308 s; Satimage数据集分别为40 s, 53 s, 9 s, 8 s; Shuttle数据集分别为458 s, 475 s, 537 s, 367 s. 数据集Letter中样本较多, 且类别数和属性数也较多, 使得训练时间较大. 虽然数据集Shuttle的类别数和属性数不多, 但由于样本较多也使得使用各种算法的训练时间较长. 数据集Satimage由于样本和类别均较少, 训练时间和测试时间也都很小. 但从实验数据上可知, 新型算法在训练时间上均优于其他3种算法.

在分类精度方面, 对于Letter数据集, 取得最好识别率的算法为one-against-one算法, 得到识别率96.6%, 这与采用新型算法得到的识别率96.5%不相上下; 对于Satimage数据集, 采用欧式距离极大极小法得到的识别率最高, 为91.7%, 这与采用新型算法得到的识别率91.6%也相差无几; 对于Shuttle数据集, 采用新型算法得到的识别率最高, 为98.4%. 因此, 该新型算法并没有降低原有算法的识别率, 有些情况下甚至可以提高识别效率. 这种情况主要是因为采用了合理的类划分方案, 能够保证先将最大的类和最易分的类先分开, 并将分开的类样本全部取出, 减少训练时参与的样本数目, 合理减少训练时间. 同时, 在类间可分性度量方面兼顾原有算法的优点, 保证算法的识别效率不降低. 然而, 与其他文献中取得的结果相比, 在数据上还有一些误差, 主要是因为在实验时采用的编程方法以及软件的架构不同. 大部分文献采用LIBSVM软件包, 而本文实验数据是基于SVM-Light软件开发包实现的. 在相同平台上对不同数据集进行实验取得的数据才具有可比性, 从上述实验结果可知, 新型算法可以在保证不降低分类精度的情况下, 大大减少训练时间.

5 结 论

本文详细研究了决策树与SVM相结合的多类分类模型, 并重点介绍了两种层次结构及其研究进展. 由于不同决策树的结构推广能力各不相同, 如何确定决策树的生成规则, 使分类器的性能有所提高也是需要考虑的问题. 在综合考虑待分类样本数以及类别的易分性能的基础上, 折衷考虑“先分样本数较大的类”和“先分易分的类”, 提出了一种基于样本的新的类划分方案, 并采用平衡决策树结构, 得到了一种新的SVM多类分类算法. 通过对大规模数据集Letter, Satimage和Shuttle的多类分类实验结果可知, 该算法不仅能大大减少训练时间, 还能得到较高的识别率, 并具有较好的推广能力, 是一种有效的多类分类算法.

参考文献(References)

[1] Vapnik V. The nature of statistical learning theory[M]. New York: Springer, 1995.

[2] 厉小润, 赵光宙, 赵辽英. 决策树支持向量机多分类器设计的向量投影法[J]. 控制与决策, 2008, 23(7): 745-750. (Li X R, Zhao G Z, Zhao L Y. Design of decision-tree-based support vector machines multi-class classifier based on vector projection[J]. Control and Decision, 2008, 23(7): 745-750.)

[3] Weston J, Watkins C. Multi-class support vector machines[R]. London: Royal Holloway University of London, 1998.

[4] KreBerl U PairWise. Classification and support vector machines[C]. Advances in Kernel Methods Supprot Vector Learning. Cambrige: MIT Press, 1999: 255-268.

[5] Bottou L, Cortes C, Denker J. Comparison of classifier methods: A case study in handwriting digit recognition[C]. Proc of the 12th IAPR Int Conf on Pattern Recognition. Jerusalerm: IEEE, 1994, 2: 77-82.

[6] Platt J C, Cristianini N, Shawe Taylor J. Large margin DAGs for multiclass classification[C]. Advances in Neural Information Processing Systems. Cambridge: Mtt Press, 2000: 547-268.

[7] 应伟, 王正欧, 安金龙. 一种基于改进的支持向量机的多类文本分类方法[J]. 计算机工程, 2006, (16): 74-76. (Ying W, Wang Z O, An J L. Study on multiclass text ctegorization method based on improved support vector machine[J]. Computer Engineering, 2006, (16): 74-76.)

(下转第156页)