一种改进的有向无环图支持向量机分类算法

王晓锋

(渤海大学 数学系, 辽宁 锦州 121013)

摘要: 针对有向无环图支持向量机多类分类方法未采用有效的有向无环图生成算法, 提出了一种改进的有向无环图生成算法。该方法采用了聚类分析中类距离的思想作为层次分类依据。实验结果表明, 该方法与原方法相比具有较高的分类精度。

关　键　词: 支持向量机; 聚类; DAG; 多类分类
中图分类号: TP391　　文献标志码: A　　文章编号: 1674-0696(2009)05-0973-03

An Improved SVM Multiclass Classification Algorithm Based on DAG

WANG Xiao-feng

(Department of Mathematics, Bohai University, Liaoning Jinzhou 121013, China)

Abstract: Aiming at the problem that the SVM multiclass classification algorithm based on DAG doesn't use an effective constructing algorithm of directed acyclic graph, an improved SVM multiclass classification algorithm based on DAG is put forward. The class distance of clustering is taken as the basis of hierarchical classification. The experiment results show that the new method has higher classification accuracy than the original algorithm does.

Key words: supporting vector machines(SVG); clustering directed acyclic graph(DAG); multiclass classification

1 引　言

支持向量机(Support Vector Machine 简称 SVM)由 Vapnik等人[1-2]提出的一种基于统计学理论的机器学习方法。支持向量机具有多方面的优点, 它较好的解决了非线性、高维数、局部极小点等问题。支持向量机本身是一种两类分类算法, 将其推广到多类分类问题更具有实际意义。当前应用较广且性能较好的支持向量机多类分类算法有 one-versus-rest[3]、one-versus-one[4]、DAGSVMS[5]以及一次性求解算法等。这些多类分类算法都存在着一些不足, 如存在着不可分的盲点、训练时间或测试时间较长, 其中 DAGSVMS分类算法在分类过程中还存在着误差累积现象, 子分类器在有向无环图中的位置对分类性能的影响较大。

2 DAGSVMS多类分类算法简介

DAGSVMS是 Platt提出的, 对于 k类问题, 在训练阶段, 生成一个有向无环图的分类模型, 包括 $k(k-1)/2$个内部节点以及 k个叶子节点, 每个内部节点都是一个两类分类器, 叶子节点为最终的类别。在决策阶段, 给定一个测试样本, 从根节点开始根据分类器的输出值决定其行走路径, 如此直到达到底层的叶子节点为止, 确定样本所属的类别。图 1为 DAGSVMS对 4类样本分类的决策过程图。

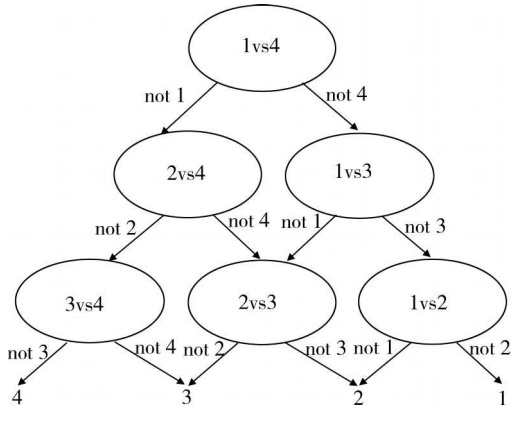

图 1　有向无环图支持向量机

DAGSVMS在决策阶段不需要遍历所有的分类器, 对测试样本判别次数为 $k-1$次, 判别次数较少, 因此相对其它的 SVM 分类算法其测试速度较快。然而, DAGSVMS在分类过程中存在误差累积现象,

收稿日期: 2009-05-15; 修订日期: 2009-06-18
基金项目: 辽宁省教育厅项目(2008Z018)
作者简介: 王晓锋(1977-), 男, 辽宁沈阳人, 讲师, 硕士, 主要从事机器学习领域的研究工作。E-mail: w200888w@163.com。

即若在某个节点处发生分类错误,则会把分类错误延续到该节点的下层节点中[6-7]。分类错误在越靠近根结点的地方发生,误差累积就越严重,分类性能就越差,所以分类器在有向无环图中的位置对分类性能影响较大。传统的DAGSVMS的有向无环图分类模型是随机生成的,并没有从减少误差累积的角度去设计有向无环图的生成算法。

3 改进的DAGSVMS多类分类算法

在DAGSVMS的分类模型中,越是上层的分类器的分类性能对分类模型的推广性能影响越大。为了减少误差累积现象,在生成有向无环图的过程中,应该让分类性能好的分类器出现在有向无环图的上层节点中。

对于含有k类样本的训练集,在生成有向无环图的过程中,每一类样本都要参与$k-1$个分类器的生成,若某类样本与训练集中的其它样本距离较远,则由该类样本生成的分类器的推广性能较好,对于这类样本参与生成的分类器,应该更早的出现在有向无环图的上层节点中。如图1,在4类样本生成的有向无环图的上两层节点中,第1类和第4类分别参与了两个分类器的生成,若此时第1类与第3、4类的距离较远,那么它们生成的分类器的推广性能较好;同理,若第4类与第1、2类距离较远,那么它们生成的分类器的推广性能也较好,说明将第1类与第4类率先生成分类器是合理的。

笔者利用聚类分析中类距离的思想作为生成有向无环图的依据,首先计算各类间的最短距离,每一个类都可得到一组与其它类的距离值,然后计算每一类与其它类距离的平均值,比较平均值的大小,平均值较大的类别,距离其它类较远,则优先选取平均距离大的类别生成有向无环图的上层分类器。

定义1:最短距离

把类A_i与类A_j中两个最近样本向量之间的欧式距离作为类A_i与类A_j之间的距离,用$d_{ij}(i,j=1,2,\cdots,k)$表示:

$$d_{ij}=\min\{\|x_a-x_b\|,x_a\in A_i,x_b\in A_j\} \quad (1)$$

显然,有$d_{ii}=0$,$d_{ij}=d_{ji}(i,j=1,2,\cdots,k)$,$x_a$为类$A_i$中的样本,$x_b$为类$A_j$中的样本。

定义2:平均距离

在含有k类的样本中,第i类与其它类的距离平均值为$w[i]$。记

$$w[i]=\frac{1}{k}\sum_{j=1}^{k}d_{ij} \quad (2)$$

则称式(2)为第i类的平均距离。

具体步骤如下:

Step 1:根据式(1)计算类与类之间的最短距离$d_{ij}(i,j=1,2\cdots,k,i\neq j)$。

Step 2:对于每个类都存在$k-1$个与其它类的最短距离值,根据式(2)计算每个类与其它类间的平均距离$w[i](i=1,2\cdots,k)$。

Step 3:比较数组w中元素的大小,对类别进行排列。当存在两个或两个以上的类别具有相同的平均距离时,首先排列编号小的类别,最后得到所有类别排列顺序为$n_1,n_2,\cdots,n_k$,$n_m\in\{1,2,\cdots,k\}$为类标号,$m=1,2,\cdots,k$(这里类别顺序$n_1,n_2,\cdots,n_k$是按照平均距离由大到小以数列的两端为起始位置交替地向中间排列,即平均距离最大的类别出现在数列的首位为n_1,次大的出现在数列的末尾为n_k,然后是$n_2,n_{k-1},n_3,n_{k-2}\cdots$,这样处于数列$n_1,n_3,\cdots,n_k$中越靠近中间位置的类别平均距离越小,越靠近数列两端的类别平均距离越大)。

Step 4:根据数列$n_1,n_3,\cdots,n_k$,生成如图2所示的有向无环图分类模型。

Step 5:算法结束。

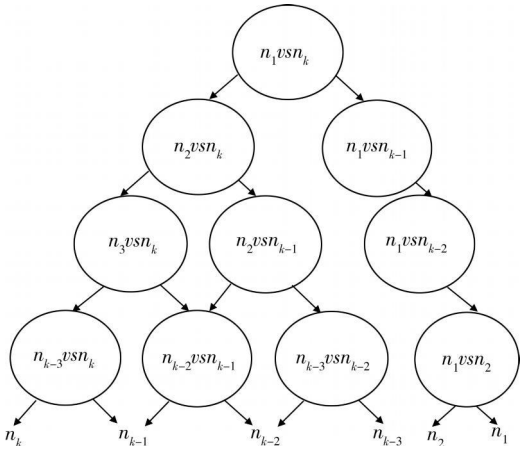

图2 有向无环图分类模型

4 实验设置及结果

实验数据如表1,数据每个特征都被归一化到[-1,+1]。所有算法均利用C++编译,在LIBSVM-2.84基础上修改得到,操作系统为Windows XP,实验平台Intel Pentium 4 Processor,512RAM。采用RBF径向基核函数$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2)$。实验中核参数$\gamma$和惩罚系数$c$的范围为:$\gamma=[2^{-15},2^{-14},\cdots,2^{0},\cdots,2^{3}]$,$c=[2^{-5},2^{-4},\cdots,2^{15}]$。由于数据集vehicle与glass没有测试数据,实验结果采用5折-交叉验证得到。将本文算法与o-v-o SVMS、DAGSVMS进行比较,表2给出了具有最好推广能力的参数和测试精度。

从实验结果可以看出,在3个数据集中中,本文算法与DAGSVMS算法相比,在分类精度上有明显提高。在letter集中中,本文算法的分类精度最高,说

表 1　实验的数据特征

测试数据	类 数	特征数	训练数据 /测试数据
vehicle	4	18	846 /–
glass	6	9	214 /–
letter	26	16	1 5000 /5 000

表 2　具有最好推广能力的参数值和测试精度

数据集		算法		
		o-v-o SVM S	DAGSVM S	本文算法
vehicle	rate	87. 35	86. 64	86. 88
	c/ γ	2^{13} /2^{-5}	2^{13} /2^{-5}	2^{13} /2^{-5}
glass	rate	72. 43	71. 50	72. 43
	c/ γ	2^{13} /2^{-3}	2^{13} /2^{-7}	2^{9} /2^{-3}
letter	rate	97. 98	97. 98	98. 00
	c/ γ	2^{5} /2^{2}	2^{5} /2^{2}	2^{5} /2^{2}

明当数据集中类别较多且数据集较大时, 本文算法具有较好的分类精度。

5 总　结

在 SVM 多类分类算法中, DAGSVM S多类分类算法具有较高的分类精度。然而 DAGSVM S分类算法中有向无环图的生成是随机产生的, 不能够很好的避免误差累积现象。笔者提出了一种新的有向无环图生成算法, 通过计算类的最短距离, 实现了分类性能较好的分类器较早的出现在有向无环图分类模型的上层节点中, 从而减少了误差累积现象带来的影响, 实验结果表明, 本文算法的分类准确率明显高于其它多类分类算法。

参考文献:

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlage, 1995.

[2] 边肇祺, 张学工. 模式识别 [M]. 北京: 清华大学出版社, 2001.

[3] Bottou L, Cortes C, Denker J. Comparison of classifier methods: a case study in handwriting digit recognition[C] //Proceedings of the 12th IAPR International Conferenceon Pattern Recognition. Jerusalem: IEEE, 1994: 77-82.

[4] KreBel U. Pairwise classification and support Vector machines[C] //Advances in Kernel Methods-Support Vector Learning. Cambrige: MIT Press, 1999: 255-268.

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

[6] 刘　勇, 全廷伟. 基于 DAG-SVMS的 SVM 多类分类方法[J]. 统计与决策, 2007(20): 146-148.

[7] 唐发明, 王仲东, 陈锦云. 一种新的二叉树多类支持向量机算法[J]. 计算机工程与应用, 2005, 41(7): 24-26.

(上接第 955页)

地名数据标准、数据交换标准、大比例尺地理空间数据质量评定标准等。

4)强化标准监督执行。充分利用重庆市测绘产品质量监督检验机构, 加强数据质量标准的监督检验, 对从事信息化测绘与地理信息数据采集、处理、生产、应用、服务、软件开发和销售等活动的企事业单位, 必须严格执行信息化测绘与地理信息标准的各项条款, 实行测绘产品执行标准通报制度。

5)建立标准一致性测试评价体系[6]。研究建立适合重庆市实际情况的一致性测试评价体系, 积极开展针对已有和在研标准以及测绘产品、设备、软件等的一致性测试和评价工作。

6)加大宣传与培训。整合全市已有标准信息资源, 畅通信息渠道, 为全社会提供及时、准确、高效、权威的标准信息服务。及时通报国内外相关标准制定、发布、实施等方面的信息, 积极主动为各类用户提供标准咨询技术服务。加大宣传贯彻力度, 扩大标准的影响, 促进标准的实施。

参考文献:

[1] 国家测绘局. 关于加快推进测绘信息化发展的若干意见[R]. 北京: 国家测绘局, 2007.

[2] 李德仁, 苗前军, 邵振峰. 信息化测绘体系的定位与框架[J]. 武汉大学学报 (信息科学版), 2007, 32(3): 189-192, 196.

[3] CJJ 100— 2004, 城市基础地理信息系统技术规范[S].

[4] 国家测绘局. 测绘标准化“十一五”规划[R]. 北京: 国家测绘局, 2006.

[5] 国家测绘局. 国家地理信息标准化“十一五”规划[R]. 北京: 国家测绘局, 2006.

[6] 刘若梅, 蒋景峰. 空间数据基础设施建设中的地理信息标准化问题[R]. 北京: 国家基础地理信息中心, 2006.