DOI:10.19358/j.issn.1674-7720.2013.12.022

基于二叉树的 SVM 多类分类的研究与改进*

周爱武,温春林,王 浩
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:支持向量机(SVM)是一种两类分类算法,如何将 SVM 算法应用于多类分类问题,目前已衍生出多种方法。其中"二叉树"方法应用比较广泛,但分类支持向量机在树中中间节点位置的不同,直接关系到该方法的分类准确性。基于二叉树方法提出了"类间相异度"的策略,根据类间相异程度来决定多类的分类顺序。

关键词:支持向量机;二叉树;超球体;相异度
中图分类号:TP312 **文献标识码**:A **文章编号**:1674-7720(2013)12-0067-03

Research and improve on multiclass classification support vector machine based on binary tree

Zhou Aiwu, Wen Chunlin, Wang Hao
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Support Vector Machine (SVM) is a binary classification algorithm, a variety of methods have derived from SVM algorithm in order to apply to multi-class classification problem. Binary tree method is used widely, but support vector machine in the different position of the intermediate nodes in the tree, is directly related to the classification accuracy of the method. The paper proposed dissimilarity between-class strategy based on binary tree method to determine the classification order, that is, according to the degree of dissimilarity between classes.

Key words: support vector machine; binary tree; super-sphere; dissimilarity

支持向量机 SVM(Support Vector Machine)[1]是一种基于统计学的 VC 维理论[2]和结构风险最小化原理基础之上的两类分类算法。目前,该算法已广泛应用于诸多领域,如人脸检验、文字/手写体识别、图像处理[3]等。支持向量机属于一种机器学习算法,核函数则是其中的核心部分。对于难以分类的低维空间向量集,通常的做法是向高维空间集转化,但这也增加了计算的复杂度,即维数灾难[4]问题。而核函数[4]却可以很好地解决这个问题,只要选取合适的核函数,即可得到高维空间的向量机(也称超平面[2])。

当使用向量机进行多类分类时,需要将多类问题转化为两类问题。常用的有"一对多"(One Versus Rest)[5]、"一对一"(One Versus One)[6]、"二叉树"(Binary Tree)[7]和"有向无环图"(Directed Acyclic Graph)[8]等方法,本文将对多类分类支持向量机[9]的这些方法作概略介绍和比较,同时对基于偏二叉树多类分类向量机提出一些改进意见。

1 SVM 多类分类方法

1.1 SVM 多类分类方法介绍

现有一个多类分类问题,其中类别数为 k。当使用支持向量机对此问题进行分类时,需假设一类为正样本,另一类为负样本。

"一对多"方法将类 i 样本作为正样本,而除该类以外的所有类作为负样本,在这两类样本间训练出向量机,该方法总共构造了 k 个分类支持向量机。在对某向量进行测试时,取计算出最大值的向量机所对应的类别作为该向量的类别。

"一对一"方法是从分类问题中选取类别 i 和类别 j 中的样本数据训练两类间的分类向量机,这样构造出的向量机的总数为 $k(k-1)/2$。虽然"一对一"分类方法产生的分类向量机的数目是"一对多"方法的 $(k-1)/2$ 倍,但"一对一"方法的训练规模要比"一对多"方法小很多。对向量的测试采取计分的方式,通过 $k(k-1)/2$ 个分类机的

* 基金项目:安徽省教育厅重点项目(KJ2009A57)

计算以后，选取得分最高的类别作为该测试数据的类别。

二叉树方法是将两类之间的 $k-1$ 个向量机作为中间节点，叶子节点对应 k 个类别样本，以这样的方式构建一棵分类二叉树，常用的方式包括满二叉树和偏二叉树。在对样本进行训练时，根节点的向量机在全部样本空间上进行训练，而子节点向量机则在根节点的负样本类或正样本类上训练，依次类推，直至 $k-1$ 个分类机在 $k-1$ 类和 k 类样本上进行训练。

有向无环图方法与“一对一”方法一样，也是在任意两类之间训练分类向量机，也即具有相同的分类向量机数目。$k(k-1)/2$ 个分类向量机作为图的中间节点，图中叶子节点为 k 类样本。但在测试向量数据所属类别时，仅需经过 $k-1$ 个分类向量机节点即可判断测试数据的类别。

1.2 基于二叉树的 SVM 多类分类方法

在 SVM 多类分类算法中，分类树是一种应用十分广泛的多类分类策略。但分类向量机在树中所处的节点位置，直接影响到分类的准确性和推广的性能。不同的二叉树结构，会使得测试数据得到不同的分类结果。随着节点分类层次的深入，可能会产生分类“误差累积”的现象[10]。因此，生成合适的二叉树结构显得异常重要。

生成多类分类二叉树通常包括两种思路：第一种是依据类中样本点的分布情况，优先分出分布区域较大的类；第二种是依据类间距离作出判断，优先分出离其他类较远的类。而衡量类分布情况的一个有效方法是计算各个类的超球体的体积，体积越大，类的分布区域也就越大。类的超球体体积定义如下：

定义 1 超球体[11]最小包含类体积：设类 S 有 n 个样本 $x_1, x_2, \cdots, x_n$，此类样本集的重心为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$，而最小包含这些样本的超球体半径为 $R=\max_{x_i \in S}\{||\bar{x}-x_i||\}$，超球体的体积为 $V=\pi R^m$，其中 m 表示样本的维数。

以偏二叉树为例，在求出各类的超球体体积后，以降序的方式排列各类的体积。将超球体体积最大的类作为正样本，除该类以外的类作为负样本，训练最优分类向量机，并将该向量机作为根节点，正样本类作为左孩子，以同样的方式在负样本中构造出最优分类向量机，并将此向量机作为根节点的右孩子节点，最终将会构造出多类分类偏二叉树。

以上根据各类超球体体积大小的方式所构造的二叉树，在很大程度上优化了分类二叉树结构。但同时也存在一些问题（图 1 所示）：虽然类 $S4$ 分布不是很广，即超球体体积比较小，但该类离其他类均较远。可见按照体积大小构造二叉树，显然是不合适的。

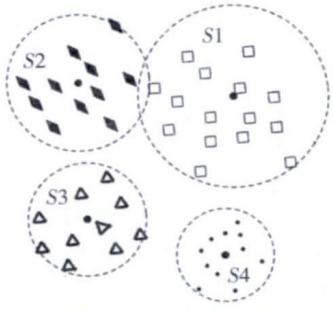

图 1 样本分布情况

本文对于构造偏二叉树提出了类间相异度的方法，有效解决了上述问题。

2 改进的偏二叉树 SVM 多类分类方法

本文从类在空间中的分布情况和类间距离这两方面着手，优化分类偏二叉树的结构。对于类的分布情况采用参考文献[11]所提出的超球体的体积来度量，而类间距离采用超球体重心间的欧氏距离来度量，关于欧氏距离的概念见定义 2。为综合考虑以上这两个方面，本文引入了类间相异度的概念，具体内容见定义 3。

定义 2 欧氏距离：也称欧几里得距离，是一个计算数据对象间距离的定义，m 维空间中两点 i 和 j 之间的真实距离表示为：$d_{i,j}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$。

定义 3 类间相异度：类与类之间的相异程度的度量，设类 $S1$ 和 $S2$ 的重心分别为 $\bar{x}_{S1}$ 和 $\bar{x}_{S2}$，$d_{S1,S2}$ 表示这两个类的重心欧氏距离，则类间相异度表示为：$D_{S1,S2}=\frac{d_{S1,S2}}{V_{S1}+V_{S2}}$，其中 V_{S1} 和 V_{S2} 分别为类 $S1$、$S2$ 的超球体体积。

为提高计算效率，本文算法均采用超球体的半径 R 来代替超球体的体积，如类 $S1$ 与类 $S2$ 的相异度表示为：$D_{S1,S2}=\frac{d_{S1,S2}}{R_{S1}+R_{S2}}$。

对于生成分类偏二叉树结构的问题，实质上是确定分类顺序问题。如何确定这个分类序列，是本文的重要内容。改进的算法采用某类与其他任一类间的相异度的总和作为该类与其他所有类间的相异度。如类 Sx 与其他所有类间的相异度表示为：$D'_{Sx}=\sum_{i=1,i\neq x}^{k} D_{Sx,Si}$。$D'_{Sx}$ 的值越大说明类 Sx 与其他所有类间的相异度越大，越要提前分割；相反，则类 Sx 与其他所有类间的相异度越小，则越要推迟分割。需要注意的是，在每确定一类分类顺序后，剩余类中任意类与其他所有类的相异度需重新计算。按类间相异度生成优先分类序列的算法描述如下：

输入：包含 n 个样本对象（含分类号）的数据集 D。

输出：包含 K 个元素的优先分类序列 S。

算法：

(1)计算每个类的最小超球体的重心和半径；

(2)repeat；

(3)根据定义 3 计算每个类相对 D 中其他剩余类的相异度之和；

(4)选择步骤(3)中相异度最大的类 i，把类标号 i 添加到 S 中，删除 D 中类标号为 i 的元素；

(5)until D 中只剩两个类的元素；

(6)把剩余的两个类的类标号添加到序列 S 中。

算法在步骤(5)返回步骤(3)循环执行，当数据集中仅包含两类样本时算法结束。

以生成分类偏二叉树的根节点和左右孩子为例，取出分类序列 S 中第一个元素的类标号，将该类和其他类间训练出的向量机作为根节点，该类作为左孩子，然后再从分类序列 S 中取出第二个元素的类标号，将该类和其他类间训练出的向量机作为右孩子。以同样的方式生成剩余的中间节点和叶子节点，最终构建出的多类分类偏二叉树如图 2 所示。

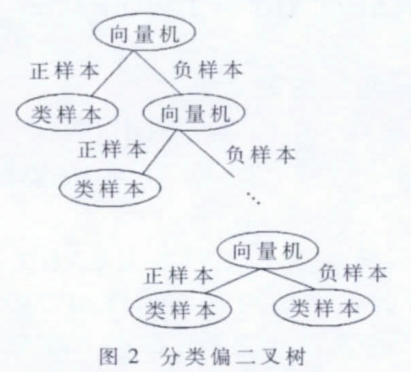

图 2 分类偏二叉树

3 实验分析

本文所有算法均使用 C++语言实现，并使用 VC6.0 完成编译。实验平台：Pentium® Dual-Core CPU 2.80 GHz、2 GB 内存、Windows XP 操作系统。所有实验数据均来自 UCI 数据库中的多类别数据集 vehicle 和 letter，具体样本数量和维数如表 1 所示。

表 1 实验数据信息表

数据集	训练样本数	测试样本数	类别数	向量维数
vehicle	600	300	4	18
letter	900	1 000	26	16

根据类间相异度的方法所生成的偏二叉树，以及仅依据类间距离和仅依据类的分布情况生成的偏二叉树，分类准确性的比较如表 2 所示。

表 2 各方法最高测试准确率所对应的参数及其准确率

数据集	依据类超球体体积		依据类间距离		依据类间相异度	
	(γ, C)	准确率/%	(γ, C)	准确率/%	(γ, C)	准确率/%
vehicle	(16,16)	65.7	(16,64)	65.4	(16,64)	66.0
letter	(2,8)	67.08	(2,8)	66.96	(2,16)	67.9

由于取不同的核参数 γ 和惩罚系数 C[12]对模型的推广有很大的影响，为了能更好地比较出依据不同的策略生成的偏二叉树的推广性能，本实验与参考文献[12]类似，对相同数据集的每一种策略均采用多种(C,γ)参数进行实验，其中 C 的取值为 2、4、8、16、32、64，γ 的取值为 2、4、8、16，这样总共有 6×4=24 种组合，每个实验的 KTT 停止条件的容许误差为 0.001。取出最高的预测准确率所对应的(C,γ)参数及其准确率进行比较。

从实验数据的分析中可以看出，对于数据集 vehicle，当训练样本的数量为 600 时，在预测准确率方面，使用本文提出的方法与其他方法相比，并没有明显的提高。而对于数据集 letter，由于比 vehicle 数据集在训练时多出 300 个样本，本文提出的方法在准确率方面有了明显的优势。总体来讲，本文提出的根据类间相异度的策略生成的偏二叉树要比单独根据类间距离或单独根据类样本的分布情况生成的偏二叉树，在准确率方面有一定的改善。

基于二叉树多类分类方法是 SVM 算法在多类分类问题中的一个重要应用，但支持向量机节点在二叉树中所处位置的不同对分类的准确性有较大影响。本文首先分析和比较了由 SVM 算法所产生的多类分类方法，然后提出了一种依据类间相异度的策略来生成基于偏二叉树的多类分类支持向量机。实验结果表明，改进的算法在准确性方面有很大的提高。

参考文献

[1] VAPNIK V N.Statistical learning theory[M].New York：John Wiley and Sons，l998.

[2] 瓦普尼克.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] 叶磊，骆兴国.支持向量机应用概述[J].电脑知识与技术，2010，6(34)：153-154.

[4] 顾亚祥，丁世飞.支持向量机研究进展[J].计算机科学，2011，38(1)：14-17.

[5] BOTTOU L，CORTES C，DENKER J.Comparision of classifier methods：a case study in handwriting digit recognitiong[C].Proceedings of International Conference on Pattern Recognition，1994：77-87.

[6] KRELEL U.Pairwise classification and support vector machines[M].Cambridge，MA：MIT Press，1999：255-268.

[7] 刘健，刘忠，熊鹰.改进的二叉树支持向量机多类分类算法研究[J].计算机工程与应用，2010，46(33)：117-120.

[8] PLATT J C，CRISTIANINI N，SHAWE-TAYLOR J.Large margin DAGs for multiclass classification[C].Advances in Neural Information Processing Systems，Cambridge，MA：MIT Press，2000：547-553.

[9] HSU C W，LIN C J.A comparsion of method for multiclass support vector machine[J].IEEE Transaction on Neural Networks，2002，13(2)：415-425.

[10] 孟媛媛，刘希玉.一种新的基于二叉树的 SVM 多类分类方法[J].计算机应用，2005，25(11)：195-196，199.

[11] 唐发明，王仲东，陈锦云.支持向量机多类分类算法研究[J].控制与决策，2005，20(7)：746-749.

[12] 单玉刚，王宏，董爽.改进的一对一支持向量机多分类算法[J].计算机工程与设计，2012，33(5)：165-169.

(收稿日期：2013-01-22)

作者简介：

周爱武，女，1965 年生，副教授，主要研究方向：数据库与 Web 技术，数据仓库与数据挖掘，信息系统安全。

温春林，男，1983 年生，硕士研究生，主要研究方向：数据库与 Web 技术，数据挖掘。

王浩，男，1987 年生，硕士研究生，主要研究方向：数据库与Web 技术，数据挖掘。