一种改进的二叉树多分支持向量机算法*

周爱武，吴国进，崔丹丹
(安徽大学 计算机科学与技术学院，安徽 合肥 230039)

摘 要：二叉树支持向量机分类算法主要是构造一个偏二叉树或是构造一颗完全二叉树，但是偏二叉树分类的准确性虽高而分类的效率低，完全二叉树分类的效率高但是准确性不高。本文提出一种算法，结合了以上两种二叉树构造方法的优点，并且更能反映样本的真实分布。实验结果表明，新算法具有较高的推广性能。

关键词：二叉树；支持向量机；多类多分；球结构

中图分类号：TP391 **文献标识码**：A **文章编号**：1674-7720(2011)04-0014-03

An improved binary tree points support vector machine (SVM) method

Zhou Aiwu, Wu Guojin, Cui Dandan
(College of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Binary tree support vector machine is mainly to construct a classification algorithm partial binary tree or tectonic a single fully binary tree. But slant binary tree classification accuracy is high but low efficiency, fully binary tree classification is high efficiency but not high accuracy. This paper proposed an algorithm that combines the above two kinds of binary tree structure method's metrics and it better reflects the real distribution. Experimental results show that the new algorithm has good generalization performance.

Key words: binary tree; support vector machine; many kinds of points; ball structure

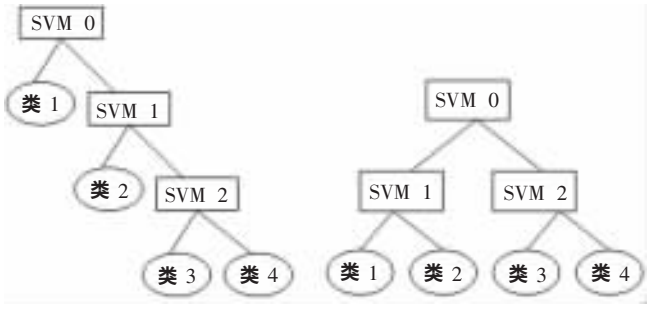
支持向量机 (SVM)[1-2] 是一种基于统计学理论的机器学习方法，由 VAPNIK 和 CORTES 于 1995 年首先提出。在解决小样本、非线性及高维向量空间的模式识别中具有良好的性能。支持向量机的思想是：如果是线性不可分，则通过某种非线性映射，将输入向量 x 映射到一高维特征空间 Z，在这个空间中构造最优超平面，把不同的类分开；如果是线性可分，就可以直接构造最优超平面。但是传统的支持向量机分类方法是针对两类问题的二分方法，而实际应用中，更多的是多类问题，如何将一个两类分类方法扩展到多类的分类方法一直是人们研究的重点，目前 SVM 多类分类方法应用得较为广泛的有："一对一"OVO (One-Versus-One)[3]、"一对多"OVR (One-Versus-Rest)[4]、"有向无环图"(DAG)[5]。这些方法都是通过构造一系列的 SVM 二分器并将它们组合起来实现的多类分类。但是都存在不足之处，前两种方法存在线性不可分区域，第三种方法虽然解决了不可分区域问题，但各子类分类器在有向无环图中的位置会影响整个分类器的性能。所以人们又提出一种利用二叉树构造 SVM 的多类分类方法。

1 BT-SVM 多类分类思想

BT-SVM 的思想是：首先将所有类别分成两子类，再将子类进一步划分成两个次级子类，如此循环下去，直到所有的节点只包含一个单独的类别为止，这些节点也是二叉树的叶子节点，这样就得到了一棵二叉树。该方法将一个多类分类问题转化为一系列的两类分类问题，其中每个子类间的分类器都是 SVM 二值分类器，对于一个 K 类问题只需要构造 K-1 个分类器，这样相对于"一对一"、"一对多"及"有向无环图"方法构造所需的分类器都要少。另外，二叉树方法可以克服传统方法遇到的不可分问题。

二叉树结构的生成：例如，对于一个四类问题，可以有图 1 中的两种二叉树结构 (还有其他的结构没有列出)。对于不同的二叉树，会得到不同的分类模型，它们的推广性能也会不同。不同的层次结构对分类精度有一定影响，并且这种影响有可能产生"误差累积"现象，既

* 基金项目：安徽省教育厅重点项目(KJ2009A57)



(a) 二叉树结构一 (b) 二叉树结构二

图 1 常用的二叉分类树

若在某个节点上发生分类错误，将会把错误延续下去，该节点的后续下一级节点上的分类就失去了意义。越是上层节点的子分类器的分类性能对整个分类模型的推广性影响越大，因此，二叉树的结构生成问题是许多学者研究的重点。目前已经有大量此类论文研究分类相同的模型。

2 几种常用的二叉树生成算法

2.1 构造偏二叉树

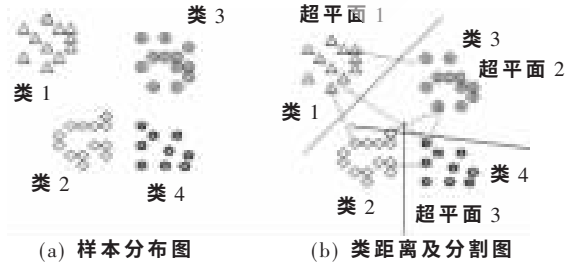
由于上层节点的 SVM 子分类器的分类性能对整个分类模型的推广性影响最大，所以在二叉树的生成过程中，应该让与其他类别相差最大的类首先分割出来。此分类的基本思想是：利用聚类分析中的类距离作为二叉树的生成算法，让与其他类距离最远的类最先分割出来。图 2 为四类样本数据的二维空间分布图，所以应该先把与其他三类距离最远的第 1 类分割出来。在剩下的三个类中，第 3 类与其他两类的距离最远，所以再把第 3 类分割出来。剩下的第 2 类与第 4 类构造最后的二值分类器。这样就得到了一棵类似于图 1(a)所示的二叉树。

定义类距离[6]：把类 S_a 与类 S_b 中两个最近样本向量之间的欧式距离作为两类之间的距离，即：

$$d_{a,b}=\min\{\| x_i-x_j \| x_i \in S_a, x_j \in S_b\} \quad (1)$$

2.2 构造完全或近似完全二叉树

如果类别个数 $N=2^e$，e 为正整数，这样就可以构造一个满二叉树，否则就构造一个完全或近似完全二叉树。该算法同样需要用到类距离，可以用式(1)定义的类距离。构造二叉树的过程如下：如果有 N 个类别，将 N 个类置于集合 S 中，S_1 和 S_2 为两个空集合，N_s、N_{s1}、N_{s2} 分别表示 S、S_1、S_2 集合中类的个数。首先，如果 $N=2$，就把其中一个类作为左子类，另一个类作为右子类，结束。如果 $N>2$，则根据类距离公式，把距离近的类放在一起，平均分成两个集合，如果 N 为偶数，每个集合中的类别个数为 $\frac{N}{2}$，如果 N 是奇数，则其中一个集合的个数为 $\lfloor \frac{N}{2} \rfloor$，另一个集合的个数为 $\lfloor \frac{N}{2} \rfloor+1$，把这两集合分别加入 S_1 和 S_2 中，把 S_1 作为二叉树的左子类，把 S_2 作为二叉树的右子类。如果 $N_{s1}=1$，则结束；如果 $N_{s1}>1$，则按照同样的方法把 S_1 分成两子类，直到子类的元素个数为 1 时结束。按照同样的方法对 S_2 进行划分。参考文献[7]中有算法的详细介绍。据此算法就可以得到一棵完全或近似完全的二叉树。



(a) 样本分布图 (b) 类距离及分割图

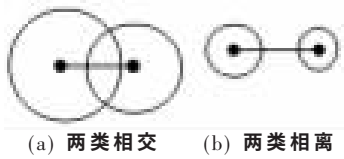
图 2 四类样本数据的二维空间分布图

上面简单介绍了两种构造二叉树的方法，现在分析这两种方法的优缺点。第一种方法每次把与其他类距离之和最大的类分割出来，这样分类的准确性较高，但是，如果对于一个 N 类分类问题，其中有一个类别 K，有可能在第一次就被分离出来，也有可能在第 $N-1$ 次才被分离出来，这样分类的效率就会比较低，而且训练时间也比较慢。第二种方法，采用完全或近似完全二叉树，所以分类和训练的效率比较高，但是，在构造这颗二叉树时，每次都会把集合中的元素平均地分成两类，这是不现实的，因为，不能保证每个集合中的任一元素相对于所属集合的相似度大于另一个集合的相似度，这样分类的准确性就会较低。分类的准确性和分类的效率是一对矛盾，本文提出一种改进算法，既考虑分离的准确性，又能保证分类的效率。

3 改进的二叉树生成算法

3.1 相似度量函数

第 2 节定义的类距离，是将两类样本的最短距离作为两类的距离，这种方法虽然简单，但是没有考虑类样本的分布情况。本文采用球结构的距离计算方法[8]，该方法在定义类距离时既考虑了类中心又考虑了类的样本分布，是一种比较科学的方法。如图 3 所示的两类，它们的类中心距离相等，但是样本分布不同。图 3(a)为两类相交，图 3(b)为两类相离。显然图 3(a)比图 3(b)具有更高的相似性。因此，不能只以类中心的欧氏距离作为相似性度量函数，还需要考虑类样本的分布情况。球结构的 SVM 能构造出半径最小且尽可能包含该类所有样本的球体，因此球体的半径可以用来度量类样本的分布。



(a) 两类相交 (b) 两类相离

图 3 两类样本的分布情况

根据上述分析，可以用下面的距离计算公式作为类 i 与 j 间的相似性度量：

$$d_{ij}=\|(a^i-a^j)\|_2-(R^i+R^j) \tag{2}$$

其中：a^i，a^j，R^i，R^j 分别是第 i 类和第 j 类的中心和半径。当 $d_{ij}\geqslant 0$ 时，说明第 i 类与第 j 类没有相交区域。d_{ij} 越大，说明第 i 类与第 j 类可分性越强，d_{ij} 越小，说明两类的相似度越高。

3.2 改进算法

定义类平均距离：

$$\bar{d}=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}d_{ij} \tag{3}$$

其中：N 为总类别数，d_{ij} 为(2)式定义的类 i 与类 j 的距离，当 $i=j$ 时，定义 $d_{ij}=0$。另定义集合 S、S_1 和 S_2，N_s、N_{s1}、N_{s2} 分别表示 S、S_1、S_2 集合中类标号的个数。算法具体描述如下：

(1)对于一个 N 类问题，首先对 N 个类进行标号，标号为 1,2,3…，N，并将这些类标号置入集合 S 中。由参考文献[9]中的式(10)和式(15)可得到每一类样本所在超球体的中心 a^i 和半径 R^i $(i=1,2,3,\cdots,N)$。根据式(2)和式(3)计算出 d_{ij} 和 $\bar{d}$。

(2) 如果 $N_s=2$，则把其中的一个类标号的类作为左子类，另一个类作为右子类，结束。

(3)计算集合 S 中的每一类到其他类的距离 d_{ij}，统计每个类到其他类距离小于 $\lambda\bar{d}$ 类的个数（λ 是一个大于零的参数，对于不同的样本分布情况，通过调整 λ 值来提高分类的精度），记作：$Num_i(i=1,2,3,\cdots,N)$。例如第 i 类，统计这个类到其他 $N-1$ 个类的距离小于 $\lambda\bar{d}$ 的类个数，计入 Num_i。找出最大的 Num_i（如果存在不唯一，任意选择其中一个），如果 $Num_i=N_s$，则转入(4)，否则，将标号 i 置入集合 S_1，并将满足 $d_{ij}<\lambda\bar{d}(j=1,2,3,\cdots,N_s,\ j\neq i)$ 的类的标号置入 S_1，将 S 剩下的类的标号置入 S_2。S_1 中类标号对应的类就作为二叉树的左子类，S_2 中类标号对应的类作为二叉树的右子类。转到(5)。

(4)如果出现这种情况，说明类之间的相似度比较高，可以根据参考文献[11]中提出的二叉树生产算法，构造一颗完全或近似完全二叉树。结束。

(5)经过步骤(3)，集合 S_1 中的类别相似度比较高，可以根据参考文献[11]中提出的二叉树生产算法，以集合 S_1 对应的左子类作为顶层节点构造一颗完全或近似完全二叉树。

(6)如果 $N_{s2}=1$，则结束。否则将集合 S_2 中的类别号置入 S，将得到的右子类作为顶层节点，回到步骤(2)。

经过上面的步骤，可以构造出一棵二叉树，这个二叉树可能是一个偏二叉树，也可能是一个完全二叉树，但是这两种都是极端的情况。更多的情况下构造的二叉树总体上是一棵偏二叉树，局部是一棵完全或近似完全二叉树。这样做的好处是，既保证了分类的准确性，又保证了分类的速度。

4 实验分析

下面以 $N=9$ 的情况分析本文提出的算法构造的二叉树，并与参考文献[11]中构造的完全二叉树做比较。图 4 是 9 类样本的球结构在二维空间分布情况，图 5 是球心坐标，图 6 是根据式(2)计算出的各类间的距离。由图 6 中的数据可以计算出 $\bar{d}=2.966$，λ 取 0.5。根据本文提出的算法，构造出图 7(a)所示的二叉树，图 7(b)图是根据参考文献[11]算法构造的二叉树(构造的偏二叉树在这里没有画出)。

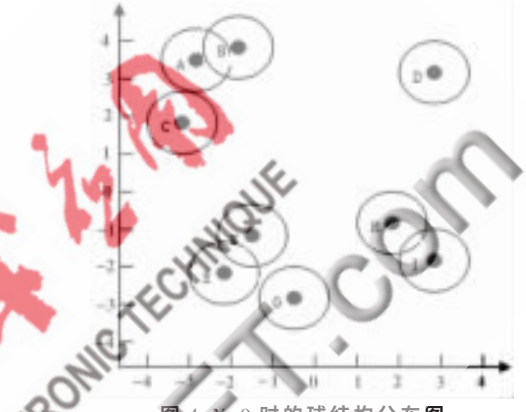


图 4 $N=9$ 时的球结构分布图

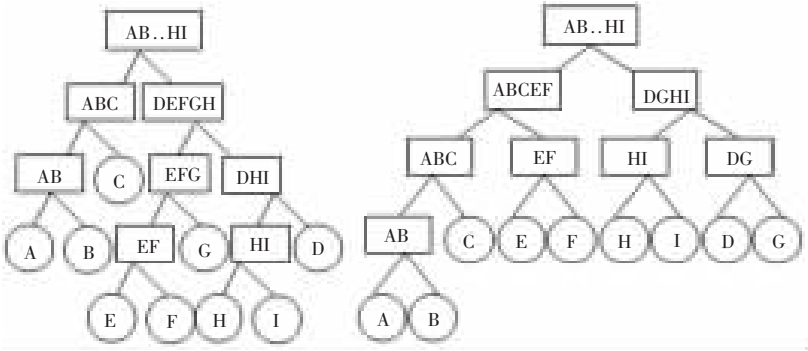
类标号	类中心坐标
A	(-2.75,3.5)
B	(-1.8,3.75)
C	(-3.25,1.75)
D	(2.95,3.1)
E	(-2.1,-2.1)
F	(-1.5,-1.2)
G	(-0.6,-2.9)
H	(2.2,-0.5)
I	(2.8,-2)

图 5 类标号对应的类中心坐标

从图 4 的样本分布图可以看出，图 7(a)所示的二叉树分类更符合样本的分布情况。而图 7(b)所示的二叉树把原本相似度非常高的 E、F、G 三个类拆分成了 EF 和 G，这显然是不合理的，出现这种情况的原因是因为此算法要求构造一个完全二叉树，左子树和右子树包含的类别个数只能相差 0 或 1，所以有些相似度高的类就被拆分开了。这样就会产生分类误差，而且本实验在一开始就出现这种误差会影响后面的

类别	B	C	D	E	F	G	H	I
A	-.0818	0.020	3.914	3.838	3.063	4.702	4.984	6.014
B		0.870	2.994	5.449	3.159	4.957	4.036	5.564
C			4.545	2.218	1.630	3.952	4.096	5.318
D				5.448	4.388	5.172	1.877	3.302
E					-0.718	-0.100	2.788	3.101
F						0.124	1.988	2.872
G							1.888	1.911
H								-0.248

图 6 各类之间的距离



(a) 二叉树 (b) 参考文献[11]算法构造的二叉树

图 7 两种分类二叉树对比

分类结果，出现误差累积的现象。而本文提出的算法，首先把相似度高的 ABC 先分割出来，再把 EFG 分割出来，最后把HI 和 D 分开，这样的结果符合样本的真实分布，所以具有比较高的分类精度。

再把图 7(a)构造的二叉树与偏二叉树作一下比较，偏二叉树每次只有一个类被分离出来，所以训练速度比较慢，而且在后面分类时效率也比较低。而本文构造的二叉树每次会把相似度比较高的一些类先分出来，再将这些类构造一个完全或近似完全二叉树，所以训练的时间会比偏二叉树低而分类的速度要更快。

本文结合偏二叉树和完全二叉树的构造思想，提出了一种基于球结构的二叉树生产算法，利用该算法构造出的二叉树更接近样本的真实分布，具有较高的分类精度和分类速度。但是算法还存在一些没有解决的问题，例如：在算法中要求 $d_{ij}<\lambda\overline{d}$，本文参数 λ 取 0.5，对于其他样本 λ 值可能会取不同的值，所以 λ 的取值问题是今后研究的重点。

参考文献

[1] VAPNIK V. Statistical learning theory[M].New York:Wiley, 1988.

[2] 邓乃扬，田英杰.支持向量机.北京：科学出版社，2009.

[3] BOTTOU L, CORTES, DENKER J. Comparison of classifier Methods:a case study in handwriting digit recognition[C]// Proceedings of the 12th IAPR International Conferenceon Pattern Recognition,Jerusalerm:IEEE, 1994.

[4] KREBERL U. Pair wise classification and support vector machines[C]//Advances In Kernel Methods-Support Vector Learnning,Cambrige:MIT Press,1999:255-268.

[5] 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.

[6] 唐发明，王仲东，陈绵云.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.

[7] 张贝贝，何中市.基于支持向量机数据描述算法的 SVM 多分类别新方法[J].计算机应用研究,2007,24(11):46-48.

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

[9] Hao Peiyi, LIN Y H. A new multi-class support vector machine with multi-sphere in the feature Space[C].Lecture Notes in Computer Science.BerLin,Heidelberg: Springer-Verlag,2007:756-765.

[10] 张晓平，杨洁明.一种新的支持向量机多类分类二叉树生成算法[J].机械工程与自动化,2007(3):1-3.

[11] 谢志强，高丽，杨静.基于球结构的完全二叉树 SVM 多类分类算法[J].计算机应用研究,2008,25(11):3268-3274.

(收稿日期：2010-11-28)

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。

吴国进，男，1986 年生，硕士生，主要研究方向：数据库与 Web 技术、数据挖掘。

崔丹丹，女，1986 年生，硕士生，主要研究方向：数据库与 Web 技术、数据仓库与数据挖崛。