多类支持向量机分类算法—DDAG①

汪　政　邵良杉 (辽宁工程技术大学 电信学院 辽宁 葫芦岛 125105)

摘　要: 随着支持向量机的发展，由最初的两类分类问题逐渐推广到多类分类问题，且其思想、算法多种多样，各有千秋。主要研究以当前比较流行的以多个二类分类器组合实现多类分类器的算法之一：DDAG。提出此算法在多类支持向量机应用分类中存在的优点和不足，并针对其不足,提出一种改进的算法思想。

关键词: 数据挖掘;支持向量机;分类算法;DDAG

Categorization Algorithms Based on M-SVMs—DDAG

WANG Zheng, SHAO Liang-Shan

(Department of Electronic and Information Engineering, Liaoning Techinical University of China, Huludao 125105, China)

Abstract: With the development of SVM, its application has been extended from binary class classification problem to multi-class classification problems. Its concepts and algorithms become various and different. The main study in this paper is DDAG, which is one of the popular algorithms, whose principle is combining the binary class classifiers with multi-class classifiers. This paper also points out the strengths and weaknesses of DDAG, when it is used in multi-class SVM text classification, and presents an improved algorithmic concept to make up for the weaknesses.

Keywords: data mining; SVM; categorization algorithms; DDAG

随着信息技术，特别是网络技术、数据库技术和高性能处理器技术的飞速发展，海量数据的收集、管理和分析变得越来越方便。包括支持向量机挖掘在内的各种数据挖掘技术，更是在一些深层次的应用中发挥了非常积极的作用[1]。

支持向量机(Support Vector Machine SVM)算法作为一种分类算法已经得到广泛的应用，它根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中，可以自动寻找哪些对分类有较好区分能力的支持向量，使得类之间的间隔最大，因为具有较高的分类效率和分类精度[2]。SVM 最初只用于两分类问题，而根据许多实际需求需要进行多分类。因此，本文将运用现比较流行的算法之——DDAG。

DAG-SVMS 是由 Platt 提出的决策导向的循环图 DDAG 导出的，是针对“一对一”SVMS 存在误分、拒分现象提出的。有向无环图(Directed Acyclic Graph DAG)是一种图中每个边都是有向的，而且不存在环的图。一个有根二元的有向无环图(Rooted Binary DAG)中的节点或者没有，或者有两条边从这个节点离开。有根二元有向无环图在分类任务中常被用来定义一类函数[3]。

1　SVM 基本原理

SVM 是 Vapnik 等人在以统计学习理论为基础,提出的一种全新的机器学习方法，它是以结构风险最小化原则为其理论基础，通过选择适当的函数子集及在该子集中的判别函数使得学习机的实际风险达到最小,保证了通过有限训练样本得到的小误差分类器对独立测试

① 收稿时间:2009-10-21;收到修改稿时间:2009-11-14

集的测试误差仍然小，得到一个具有最优分类能力和推广泛化能力的学习机。

SVM 的主要思想可以概括为两点:(1)它是针对线性可分情况进行分析,对于线性不可分的情况,通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分,从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能；(2)它基于结构风险最小化理论之上在特征空间中建构最优分割超平面,使得学习器得到全局最优化,并且在整个样本空间的期望风险以某个概率满足一定上界[4]。

2　DDAG方法

(Decision Directed Acyclic Graph DDAG)是一种图中每个边都是有向，而且不存在环的图。一个有根二元有向无环图(Rooted Binary DAG)中的节点或者没有，或者有两条边从这个节点离开。有根二元有向无环图在分类任务中常被用来定义一类函数[5]。利用有根二元有向无环图计算的函数定义如下：

首先为方便研究该算法，我们可建立一个足够大的用于分类器训练的语料库，要求带有必须的标注信息。同时不失一般性，我们将 K 类分类问题的类别集合表示为{1,2,⋯,k}。COPi，i∈{1,2,⋯,k }指的是语料库中属于第 i 类的那部分语料的集合。

为了解决一个 k 类的分类问题，决策有向无环图(DDAG)必须经过下面这三个步骤：

① 随机产生一个类别序列$[t_1,t_2,\cdots,t_k]$，其中 $t_i\in\{1,2,3\cdots,k\}$。同时，给定两个下标向量，i 与 j，并将这两个变量分别初始化为指向类别序列的开始与结尾。

② 针对当前待分类的测试实例 Xtest，我们利用语料库 COP_i 和 COP_j 训练得到二类分类器 f_{ij},得到的分类器以最大边界 COP_i 和 COP_j 中的训练实例做正确分类。利用分类器对测试实例进行分类。如果 sign|fij(xk)|=1 则将 i 递增 1，否则将 j 减 1。重复这一步，直到 i 等于 j。[注：变量 xk 表示测试实例；函数 fij()表示经训练得到的二类分类器；函数 sign| |表示对于测试实例进行训练后，得到的结果，其值域为｛0，1｝。]

③ 当 i 等于 j，则针对测试实例 Xtest 的分类过程结束，将 j 最终所指向的类别作为分类器赋予 Xtest 的类别。

『训练过程：在训练分类器 f_{ij}，我们将第 i 类的训练数据 cop_i 作为正训练数据，而第 j 类的训练数据 cop_j，作为负训练数据。如此训练得到的二类分类器 f_{ij} 可以将 cop_i 与 cop_j 作正确的分类。也就是说,对 $Xtrain\in cop_icop_j$，如果 $Xtrain\in cop_i$,，也就是说实例 Xtrain 于第 i 类，那么 $sign|f_{ij}(Xtrain)|=1$，否则 $Xtrain\in cop_j$, $sign|f_{ij}(Xtrain)|=-1$。』

譬如：当 K=5 时，可用图 1 来表示出：

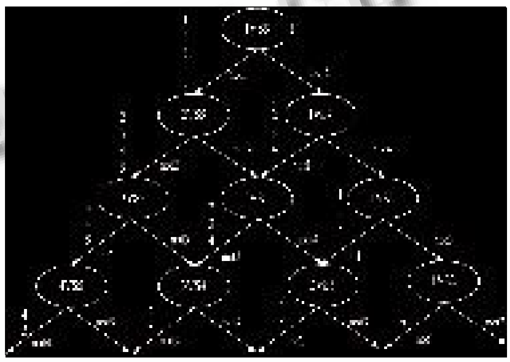

图 1　DAG-SVMS 分类结构图

从以上的叙述中，可明白决策有向无环图等价于在一个类别列表上的连续操作，列表中的每个节点对应多分类问题类别集合中的一个类别。列表用类别集合中的所有类别以随机方法做初始化。以当前列表下标(开始时指向列表的首与尾)所指向的数据训练出二类分类器，对测试实例进行分类。如果分类的结果选择了其中一个，另一个类别将被排除。这个选择的过程只需要同时修改列表的下标，得到新的列表。再次用列表首尾所对应数据训练分类器。如此循环反复，直到最终只剩下一个类别。最后剩下的类别将作为多类分类器赋予待分类实例的类别。我们容易得到，对一个 N 类问题的分类过程，决策有向无环图为了对某待测试数据库进行分类，需要构造 n-1 个决策点，也就是构建 n-1 个二类分类器。这是一种计算复杂度相当高的算法。

3　改进DAG-SVMS算法

DAG-SVMS 简单易行，只需使用 n-1 个决策函数既可得出结果，较其他方法提高了测试速度，而且不存在误分、拒分区域；另外，由于其特殊的结构，故有一定的容错性，分类精度较一般的有向无环图方

法高。

传统的 DAG-SVMS 的特点是：①根的选择是随机的；②决策走向依据决策面的值随机选取下一决策面。然而，自上而下的“误差累计”的现象是层次结构的特有弊端，DAG-SVMS 也没有能够幸免。即若在某个结点上发生了分类的错误，则会把该分类的错误一直延续到该结点的子、孙结点上。所以当分类错误在越上层的地方发生，由于其累计的效应，其分类性能越低，更严重的是在根结点发生分类错误，其分类性能会受到严重的影响[6]。

从前面的分析可以看出越靠上层节点的分类性能对整个分类模型的推广性影响越大。因此，在生成有向无环图的过程中，应该让最易分离的两类最早分离出来。有鉴于此，本文提出了根据训练样本在属性空间的几何分布情况来生成有向无环图的方法，从而建立一个推广性高的多类 SVM 分类模型。

其改进算法的基本思想为：以类样本分布情况作为有向无环图的生成算法，从而构造推广能力好的基于有向无环图的 SVM 多类分类模型。其主旨就是在每次生成有向无环图内节点时，选择最易分离的情况来构造当前节点的二值 SVM。分离顺序不一样，每个类的分离情况是不同的。最大的类和最小的类比较集中，最容易先把它们分离出来。因为各类数据的真实分布无法得知，所以用有限样本数据的分布来对真实分布作近似估计。样本分布范围的度量可采用超球体最小类包含，超球体的体积就是此类样本的分布度量。

图 2　4 类样本数据的二维输入空间分布图

图 2 为 4 类样本数据的二维输入空间分布图，直观上看，最好的分离顺序为：以第 1 类和第 4 类首先与其他类构造分离超平面，即先把体积最大的和最小的分离出去，接着再在第二层次中把最大最小的类第 2 类第 3 类和第 2 类第 4 类分离出来，第三层每个节点就有两类，直接分离就是了。这主要是考虑到各类样本空间分布范围的大小，见图 3。

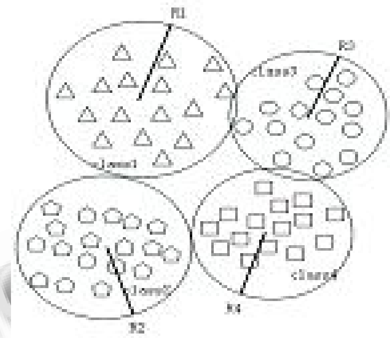

图 3　四类样本空间分布范围的大小

定义(超球体最小类包含). 设类 S 有 n 个样车 $x_1, x_2, \cdots, x_n$，此类样本集的重心为=，则最小包含这些样本的超球体半径为：

$$R = \{|| - xi ||\} \tag{1}$$

式中||　||表示欧氏距离运算，超球体的体积计算为 V=Rd，因为目的只是比较各个类别分布范围的相对大小，所以为了简化计算，可用半径代替体积，只需计算式(1)即可。

具体的算法流程如下：

Step1：根据式(1)计算各类样本数据的分布体积 $V_i(i=1,2,3,\cdots,k)$；

Step2：根据各类的分布体积由大到小的顺序，对类别进行排序。当存在两个或两个以上的类别具有相同分布体积时(这种情况很少出现)，把类标号小的类排在前面。最后得到所有类别排列 $n_1, n_2, \cdots, n_k$，此处 $n_m \in \{1,2,\cdots,k\}$，$m=1,2,\cdots,k$ 为类标号；

Step3：利用二值分类的 SVM 训练算法构造决策导向的循环图各内节点的最优超平面。在根节点处. 从样本分别选择第 n_1 类和第 n_k 样本为正样本集，其他样本为负样本集，利用 SVM 训练算法构造最优超平面，然后把属于第 n_1 类和第 n_k 样本的样本从样本集中删除。在第 2 层节点处，从样本选择第 n_2 类和第 n_k 类，第 n_1 类和第 n_{k-1} 类样本为正样本集，其他剩余的样本为负样本集，利用 SVM 训练算法构造最优超平面，然后把属于这些类的样本从样本集中删除。依次下去，最终可得到如图 4 所示的基于有向无环图的

多类别 SVM 分类模型；

Step4：算法结束，如图 4 所示[7]。

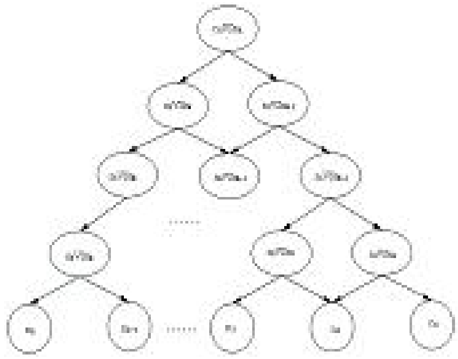

图 4 算法流程

4 实验测试

实验采用局部核函数中的高斯径向核函数，并选定了固定的参数，在 UCI 数据库的 Organic solvents(有机溶剂)数据里进行多次实验．将 DDG-SVMS 与改进后的 DDG-SVMS 进行比较,结果如表 1 所示。

表 1 DDG-SVMS 与改进后的 DDG-SVMS 进行比较

数 据	参 数	分类精度（%）	
		DAG-SVMS	Improved DAG-SVMS
Organic solvents (有机溶剂)	3	96.54	98.33

从实验的结果可以得知，改进后的算法与原来的分类器相比，分类精度大大提高。在对实验后的结果进行分析统计情况来说，改进后的算法在实际的应用中的分类准确性有了显著的提高。

5 总结和展望

实际应用中的多分类需求越来越多，随着支持向量机在这相关领域的深入运用，如何有效地将该方法推广到多类分类问题中必将引起人们广泛的兴趣。该文讨论了支持向量机的基本原理并分析了在传统 DAG-SVMS 基础在之上的一种有效的支持向量机多类分类算法。当然该算法的执行效率仍受到模型拓扑结构的制约，这也为进一步改进算法提供了有效途径，更进一步的研究还有待研究者们的继续努力[8]。

参考文献

1 邵良杉,王军.基于关联规则的手写体汉字识别技术.计算机应用, 2004,24(9):37－40.

2 衣治安,刘杨.基于二叉树的多分类SVM算法在电子邮件过滤中的应用.计算机应用,2007,27(11):2860－2862.

3 那健,丁大伟,丁俊松.支持向量机的多类分类算法.科技咨询导报,2007,29:230－231.

4 朱慕华,等.支持向量机在文本分类中的应用[硕士学位论文].沈阳:东北大学,2006.

5 张苗,张德贤.多类支持向量机文本分类方法.计算机技术与发展,2008,18(3):139－142.

6 高剑,罗志增.支持向量机在肌电信号模式识别中的应用.传感技术学报,2007,20(2):366－369.

7 刘勇,全廷伟.基于 DAG-SVMS 的 SVM 多类分类方法.统计与决策,2007,248(20):145－148.

8 朱小娟,陈特放.基于SVM的词频统计中文分词研究.微计算机信息，2007,23(30):205－207.