一种改进的 DDAGSVM 多类分类方法①

熊忠阳 陈 玲 张玉芳(重庆大学 计算机学院 重庆 400044)

摘 要: 支持向量机最初是针对两类分类问题提出的,如何有效地将其推广到多类分类问题仍是一项有待研究的课题。本文介绍了现有的具有代表性的多类支持向量机分类算法,并在分析决策导向非循环图支持向量机分类器生成顺序随机化的基础上,引入类内的分散度,以基于样本分布的类间分离程度作为类别的划分顺序,最终构成了一种分类间隔较大的决策导向非循环图支持向量机分类算法。实验结果表明了本文方法具有更高的分类精度。

关键词: 支持向量机; 决策导向无环图; 类内分散度; 类间分离程度

Improved DDAGSVM Multi-Class Classification

XIONG Zhong-Yang, CHEN Ling, ZHANG Yu-Fang

(Department of Computer, Chongqing University, Chongqing 400044, China)

Abstract: support vector machine is originally designed for binary classification. How to effectively extend it for multi-category classification is still an on-going research issue. This paper presents a general overview of existing representative methods for multi-category support vector machines. The processes of making decisions on the decision directed acyclic graph support vector machines were random. For this reason this paper inducts an internal-class degree of dispersion. An external-class separate measure is defined based on the distribution of the training samples to form the classes' separating sequences. An improved algorithm having greater classification distance for decision directed acyclic graph support vector machines is proposed. The experimental results show that it has higher multi-class classification accuracy than the original decision directed acyclic graph multi-class support vector machines.

Keywords: support vector machine; decision directed acyclic graph; internal-class degree of dispersion; external-class separate measure

1 多类支持向量机分类方法

支持向量机[1]是 20 世纪 90 年代中期在统计学习理论[2]基础上发展起来的一种新型机器学习方法。它在解决小样本,非线性及高维模式识别问题上表现出许多其它机器学习方法不可比拟的优势。支持向量机最初是针对二类分类问题提出的,但在实际应用中往往是多类分类问题,因此,将支持向量机推广到多类分类成为目前 SVM 研究的热点问题之一。当前已经有许多算法将 SVM 推广到多类分类问题,这些算法统称为“多类支持向量机”。它们可以大致分为两大类:

(1) 通过某种方式构造一系列的两类分类器并将它们组合在一起来实现多类分类;

(2) 将多个分类面的参数求解合并到一个最优化问题中,通过求解该最优化问题“一次性”地实现多类分类[3,4]。

第二类方法尽管看起来简洁,但是在最优化问题求解过程中的变量远远多于第一类方法,训练速度不及第一类方法,而且在分类精度上也不占优。当训练样本数非常大时,这一问题更加突出。正因如此,第一类方法更为常用[5]。

① 基金项目:中国博士后科学基金(20070420711); 重庆市科委自然科学基金(2007BB2372)
收稿时间:2010-04-08;收到修改稿时间:2010-05-02

下面简单介绍一些常用的实现支持向量机的多类别分类[6]。假定多类分类问题有 K 个类别 S = { 1, 2, …, k} ,训练样本为{ (x_m , y_m) , m = 1, 2, …, l} ,其中y_m ∈S。

1.1 1-v-r SVMs

1-v-r 方法(One-versus-the-rest Method)[7]构造 k 个支持向量机子分类器。该方法依次用一个两类分类器将每一类与其它所有类别区分开来，得到 K 个分类函数。测试时,对测试数据分别计算各个分类器的决策函数值,并选取函数值最大对应的类别为测试数据的类别。

1.2 1-v-1 SVMs

1-v-1 方法(One-versus-one Method)[7]是由 Knerr 提出的,该算法在每两类间训练一个分类器，因此共构造 k(k-1)/2 个 SVM 子分类器。测试时,将测试数据对 k(k - 1) /2 个 SVM 子分类器分别进行判断,并为相应的类别“投上一票”，最后选择得票最高的类别作为测试数据的类别。

1.3 DDAG SVMs

DDAGSVMs 方法是 Plantt 等提出的决策导向非循环图(Decision Directed Acyclic Graph, DDAG)[8]方法,在训练阶段与 1-v-1 方法相同,构造 k (k - 1) /2 个两类分类器;然而在决策阶段,该方法将 k (k - 1) /2 个分类器构造成一种两向有向无环图(如图 1 所示)：包括 k (k- 1) /2 个内部结点以及 k 个叶子结点,其中每个内部结点为一个两类分类器,叶子结点为最终的类值。当对一个未知样本进行分类时,从根结点开始根据分类器的输出值决定其走左结点或右结点,直到达到叶子结点为止,该叶所代表的类别即为该样本所属的类值。其优点是推广误差只取决于类数 k 和结点上的类间间隙,而与输入空间的维数无关。

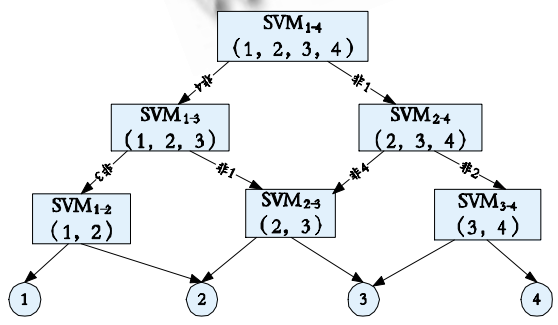

图 1　DDAGSVMs 方法

2　一种改进的DDAGSVM多类分类算法

对未知测试样本进行分类识别，是从决策树顶到达叶结点的计算和判断过程。传统的 DDAGSVM 方法:从决策树顶为 $SVM_{1/k}$ 开始，对待分类样本 x，计算决策函数值，如果非 1，则下一结点为 2/k,否则下一结点为 1/k-1。即某中间结点为 i/j 时，根据获得的结果，决定下一结点(i+1/j,或者 i/j-1)。重复该过程，直到到达树的叶结点，则待分类样本 x 就属于该叶结点所代表的类。

2.1 类间分离程度

如何估计各类间的易分性?根据训练数据估计各类间易分性，通常的做法是用类中心间的 Euclidean 距离或马氏距离作为分离性测度，但这种方法的缺点在于没有考虑训练样本在属性空间的几何分布情况。如果不考虑待分类的分布，类中心间的距离远近并不能总是正确代表类间的分离度，类分布是影响类间分离性测度的重要因素。

定义 1. $E_i = \frac{1}{n_i}\sum_{d=1}^{n_i} \Phi(x_d)$ 表示特征空间中第 i 类的

样本中心。

定义 2. $D_{ij} = \|E_i - E_j\|$ 表示第 i 类与第 j 类样本的中心距离。

定义 3. 标准差 $s_i = \sqrt{\frac{\sum_{d=1}^{n_i}(x_d - E_i)^2}{n_i - 1}}$ 表示

第 i 类样本的类内分散度。

标准差是一组数值自平均值分散开来的程度的一种测量观念。一个较大的标准差，代表大部分的数值和其平均值之间差异较大；一个较小的标准差，代表这些数值较接近平均值。σ 越小，分布越集中在样本中心附近，σ 越大，分布越分散。

定义 4. $S_{ij} = \frac{s_i + s_j - D_{ij}}{D_{ij}}$ 表示类间不可分离程度。

以类 i 与类 j 之间的相交程度来刻画两类的分离程度，$S_{ij} > 0$ 表示两类别相交，$S_{ij} \le 0$ 表示两类别相离。S_{ij} 越大，类间分离程度越低。

2.2 改进的 DDAGSVM 生成顺序

传统 DDAGSVM 方法的缺点：根的选择以及每个结点的选择性固定化，而结点的选择与分类器的性能密切相关。由于在某个结点上发生分类错误，则会把错误延续到该结点的后续结点上。分类错误在越靠近根的地方发生，由于误差的累积效应，分类性能就越差，尤其在根结点上发生分类错误，将严重影响分类性能[9-11]。

改进的 DDAGSVM 关键问题就是设计一个合理的根节点和分支结构，即每一个结点都是选择最容易分开的两个子类，本文以类间分离程度为权值构造 DDAGSVM 决策多类分类，具体过程如下：

第一步：根据定义 4 计算类别集合 S 中每两类样本在特征空间中的不可分离程度；

第二步：S_{ij} 选择最小的两类作为根结点；

第三步:假定当前结点为 $SVM_{i/j}$ 时，根据 i,j 两类 SVM 的分类函数 $f_{i,j}(x)=sign(\sum_{\text{支持向量}} y_m a_m (x\cdot x_m)-b)$ 的值，若 $f_{i,j}(x)=+1$ 表示 x 不属于类 j，那么类别集合 S=S\j ；若 $f_{i,j}(x)=-1$ 表示 x 不属于类 i，那么类别集合 S=S\i。再从 S 中选择 S_{ij} 最小的两类作为下一结点。重复该过程，直到到达树的叶结点，则待分类样本 x 就属于该叶结点所代表的类。

3 实验分析

本实验采用 VC 环境下 C++语言编程，在 libsvm 基础上修改得到。为验证改进的 DDAGSVM 决策算法的有效性，在标准 UCI 数据库的 letter, satimage, shuttle 三个数据集上进行多次实验，数据集的属性如表 1 所示。

表 1 本文实验数据集统计表

数据集	训练样本数	测试样本数	类别数	属性数
Letter	15000	5000	26	16
Satimage	4435	2000	6	36
Shuttle	43500	14500	7	9

训练 SVM 均使用高斯径向基 RBF 核函数 $K(x,y)=\exp(-\frac{\|x-y\|^2}{s^2})$，参数均采用 libsvm 默认参数。实验结果采用查全率与查准率相结合的方法进行评价。分类器在类别 C_i 上的查准率(Precision)计算公式如下：

$$Precision_i=\frac{TP_i}{TP_i+FP_i}$$

分类器在类别 C_i 上的查全率(Recall)计算公式如下：

$$Recall_i=\frac{TP_i}{TP_i+FN_i}$$

其中 TP_i 表示测试文档集中本来属于类别 C_i 而且被分类器分类到类别 C_i 的文档数，FP_i 表示测试文档集中本来不属于类别 C_i 但却被分类器错误分类到类别 C_i 的文档数，FN_i 表示本来应该属于类别 C_i 但被分类器分类到别的类别的文档数。

表 2 传统方法与改进方法的比较 %

方法	Letter		Satimage		Shuttle	
	查全率	查准率	查全率	查准率	查全率	查准率
改进前	90.07	90.15	90.32	91.03	90.12	90.45
改进后	90.62	90.52	91.72	92.87	91.05	91.01

4 结论

从实验结果可以看出，本文给出的改进 DDAGSVM 算法对提高分类正确率是有效的。通过引入类内分散度，改进的 DDAGSVM 分类决策算法克服了传统 DDAGSVM 算法在决策过程中的随机性，而是优先选择最容易分开的两个子类来构造 DDAGSVM 决策树，通过类间分离程度确定决策树的生成过程，从而最大程度的减少积累误差，提高推广性能。

(下转第 33 页)

(上接第221页)

参考文献

1 张学工.关于统计学习理论与支持向量机.自动化学报, 2000,26(1):32－42.

2 Vapnik V. The nature of statistical learning theory. pringer-Verlag, NewYork. NY,1995.张学工译, 统计学习理论的本质,清华大学出版社, 2000.

3 Weston J, Watkins C. Multi-class Support Vector Machines.Technical Report CSD_TR_98_04,Royal Holloway, University of London, Department of Computer Science, 1998.

4 Crammer K, Singer Y. Ultraconservative Online Algorithms for Multi-class Problems. Proc of the 14th Annual Conf on Computational Learning Theory, 2001.

5 余辉,赵晖.支持向量机多类分类算法新研究.计算机工程与应用, 2008,44(7):185－189.

6 黄剑锋,刘付显,朱法顺. 基于多类分类支持向量机的空袭目标识别. 微计算机信息, 2008,(10):258－260.

7 Hsu CW, Lin CJ. A Comparison of Methods for Multi-class Support Vector Machines. IEEE Transaction on Neural Networks, 2002,(13):415－425.

8 Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DDAGs for Multi-class Classification. Advances in Neural Information Processing Systems, MIT Press, 2000;12:547－553.

9 孟媛媛, 刘希玉.一种新的基于二叉树的 SVM 多类分类方法.计算机应用, 2005,25(11):2653－2657.

10 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):139－141.

11 刘洋,张秋余.基于 Huffman 树的多类 SVM 方法.计算机工程与设计, 2008,29(7):1792－1793.