DOI:10.11721/cqnuj20130621

基于二叉树结构双优化的 SVM 多分类算法研究*

徐国浪[1],魏 延[2]
(1. 重庆师范大学 数学学院;2. 重庆师范大学 计算机与信息科学学院,重庆 401331)

摘要:针对传统二叉树在多分类问题上存在分类精度不够高和时间复杂度较高的不足,提出了一种基于二叉树结构双优化的 SVM 多分类学习算法。此算法利用遗传算法对已经提取的特征参数子集和核参数进行双重优化,以获得最优的主要特征参数,从而有效地解决了样本结构复杂、分布不平坦的多分类识别问题。作者运用 UCI 数据库中的数据,通过仿真实验,并就经度和时间复杂度与有向无环图法和一对一法作比较,结果表明本文提出的算法具有较好的优越性。

关键词:GA;SVM;二叉树;多分类识别
中图分类号:TP38 **文献标志码**:A **文章编号**:1672-6693(2013)06-0109-05

传统的统计学从本质上说主要研究的是当样本趋向无穷大时的渐进理论。然而在实际生活中,样本数目通常是有限的,因此在研究小样本数据前提下的统计学习规律是一个非常有实用价值的问题[1]。然而,目前智能多分类面临的一个难题之一就是样本特征知识的发现问题:要求在不解体、实时多分类的情况下,获取大量有效样例就显得更为困难。而常规的分类方法大多是依赖于大样本情况下的统计特性,当训练样本有限时,难以保证有较好的分类泛化性。由 Vapnink 等人提出的支持向量机对小样本数据具有良好的泛化性和较好的分类精度,能够有效地解决过学习问题,恰好弥补了常规诊断的不足。基于统计学习理论的支持向量机(Support vector machine ,SVM)是在结构风险最小化原则的基础之上发展起来的通用学习方法。该方法通过将输入空间映射为高维特征空间,从而有效地降低待求解问题的 VC 维,寻求经验风险和置信范围的最小化。因而被广泛地被应用在模式识别、函数逼近、数据挖掘、故障诊断等领域[2]。

由于标准 SVM 一般解决的是两类别分类问题,而实际需要解决的一般是多分类问题。因此如何将 SVM 应用于多分类问题,对挖掘 SVM 的应用潜力将具有非常重要的意义。目前,如何运用 SVM 处理多分类问题是当前研究的热点之一。研究者们已提出一些卓有成效的多类支持向量机方法,可以归纳为 2 大类[3]:第一大类是直接的方法。该方法是 Weston 在 1998 年提出的多类分类算法为代表,只使用一个 SVM 判别函数实现多分类输出。这种算法的目标函数十分复杂,变量数目过多,计算复杂性也非常高,分类精度不理想;第二大类是组合的方法。这一大类方法是通过组合多个二值分类器来实现对多类分类器的构造。常见的构造方法是:一对一方法和一对多方法,以及在这 2 种方法基础上的改进算法,如纠错编码支持向量机、层次支持向量机、有向无环图支持向量机和二叉树支持向量机等。

本文针对 SVM 算法在解决多分类问题上存在的不足和传统二叉树对多分类问题分类精度不够高的缺陷,提出了一种基于二叉树结构双优化的 SVM 多分类学习算法。此方法首先对实际生活中特征参数运用遗传算法进行优化,然后构建基于二叉树结构的 SVM 多分类器,并对多分类器中 SVM 的核参数(C、σ^2)运用遗传算法再次优化,以获得最优分类参数。最后,在 MATLAB 实验平台上,用 UCI 数据库中的实验数据进行仿真,并取得良好效果。

1 运用 GA 算法对特征参数优化

特征选择的优劣将直接影响模式分类的效率和准确率。为了提高分类识别率,需从实验数据中提取大量的原始特征。然而从提取方法上看,许多特征不仅相互关联,而且冗余,这将直接影响识别的准确性和速度,所以

* **收稿日期**:2012-12-21 **修回日期**:2013-02-21 **网络出版时间**:2013-11-20 14:46
资助项目:重庆师范大学博士研究基金(No. 11XLB047)
作者简介:徐国浪,男,硕士研究生,研究方向为机器学习与智能计算,E-mail: xgl-120@126.com;**通讯作者**:魏延,E-mail: weiyancq@126.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20131120.1446.201306.109_048.html

要对原始数据进行选择,去掉那些相关性强或者不易判别的特征(即特征选择)。特征选择实质上是一种优化组合问题。目前优化组合的方法很多,本文选择的是遗传算法,它易于跳出局部次优解,而且无需建立优化方程,具有良好的隐并行性和稳健性,已成为信息科学、计算机科学和人工智能等学科所关注的焦点,本文利用遗传算法对样本特征进行优化选择,涉及的关键技术如下[5]:

1)个体编码。在特征选择问题中,个体的编码方式采用二进制,表达简单,操作方便,可代表较广范围的不同信息。它的长度为原始特征数,第 i 位代表第 i 个特征,值为"1"时,表示第 i 个特征项被选用;反之,为"0"时,表示该特征项未被选用。

2)评估函数(适应度函数)的确定。特征选择的目的是为了寻找出分类能最强的特征组合,因此需要一个定量准则来度量特征组合的分类能力。

$$J(X)=S_b-S_w \tag{1}$$

其中 S_b 为类模糊距离,S_w 为类内模糊距离。它们的计算方法如下(采用两模式之间模糊距离为海明距离)

$$\rho(A,B)=1-\frac{1}{n}\sum_{i=1}^{n}|\mu_A(X_i)-\mu_B(X_i)| \tag{2}$$

在计算类间距离时,$\mu_A(X_i)$、$\mu_B(X_i)$代表类 A 和类 B 的均值向量,其中均值向量可通过(3)式求出。

$$c_i=\frac{1}{n}\sum_{x\in W_i}X,\ i=1,2,\cdots,k \tag{3}$$

其中,$W_1,W_2,\cdots,W_k$ 为 k 个类别,c_i 为第 i 个类别的类中心特征向量,在 W_i 类别中有 n_i 个数据。

对任意两类均要用(2)式计算类间距离,然后相加即得 S_b。计算类内距离时,$\mu_A(X_i)$、$\mu_B(X_i)$是同一类内的数据 A 和数据 B 的特征向量。对每一类内的各数据间均要计算类内距离,然后各类的类内距离相加即得 S_W。

3)遗传操作。选择是符合自然界中优胜劣汰的过程。交叉用来指导搜索向具有潜在更好的解的区域发展(本文使用一致交叉方法)。变异是防止选择和交叉运算遗漏信息的遗传操作(本文采用变化变异率的方式)。变异率为

$$P_c=\begin{cases}k_1\cdot\exp\left(-\gamma\cdot\dfrac{f}{f_{\max}}\right)\cdot\left(1-\dfrac{R}{R_{\max}}\right), & \text{if } P_c>0.01\\ k_2, & \text{if } P_c\leqslant 0.01\end{cases} \tag{4}$$

式中,k_1 和 k_2 为 [0.001,0.01]之间的常数,而且取 $k_1>k_2$,f 为平均适应度函数值,$f_{\max}$为最大适应度值,R 为父代间的海明距离,$R_{\max}$为父代之间的最大海明距离,γ 为常数。本文假设 $k_1=0.015$,$k_2=0.002$,$\gamma=10.0$。

2 基于二叉树结构的 SVM 分类器

由 Vapnik 等人基于统计学习理论的 VC 维理论和结构风险最小化原理的基础上发展起来的 SVM 在小样本学习方面具有独特优势,目前已经在多分类问题以及状态预测方面得到了一些应用。SVM 最初是以二值分类问题为背景而提出的。多值分类(k 分类,$k>2$)问题也是以二值分类为基础的。文献[1]提出了一种带优先级的二叉树多分类故障诊断算法(2PTMC)就是 SVM 在多分类问题上的一种成功应用。这种算法首先按故障发生的概率大小排序,并构成一个集合 $P=\{P_1,P_2,\cdots,P_i,\cdots,P_k\}$($P_1$ 表示最可能发生故障,P_k 表示发生故障可能性最小一级)。然后,根据 P 建立基于 SVM 的二叉树多故障分类器,如图 1 所示。

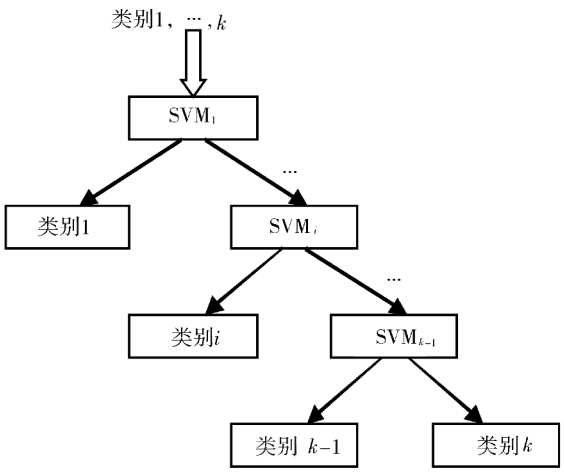

图 1 由 $k-1$ 个分 SVM 构成的二叉树分类器

这种算法把 k 类中的 $k-1$ 类看作一大类,把余下的一类看着另外一大类,建立一个二值分类器,然后再在那 $k-1$ 类中,取出$(k-1)-1$ 类来看作一大类,把那 $k-1$ 类中余下的一类看作另外一大类,建立一个二值分类器。由根节点开始对该对象的特征参数代入分类函数逐渐测试其值,当测试值为 1 时,即到达叶结点(终止前进,表示该分类工作正处于叶结点所代表的类别);当测试值为 -1 时,则顺着分支向下走,直至到达某个叶结点。依此类推,直到构建到最后一个二

值分类器才结束。对于 k 类问题，一共需要建造 $k-1$ 个二值分类器，在复杂度上，这种算法要比单纯的一对多、一对一和有向无环图法简单直观，重复训练样本量也少，训练和识别速度可以提高。证明见文献[1]。

3 基于遗传算法的核参数优化

遗传算法是通过对种群结构的迭代操作，每一次产生一组解（基因串，如二进制编码）在开始搜索时随机产生一组解，求解过程中使群体不断优化，进而找到最优解或次优解[4]。本文采用径向基核函数（RBF）作为 SVM 的核函数。由于 RBF 有两个模型参数 C（误差惩罚因子）和 σ^2（核函数的参数），因此，要对这两参数（C、σ^2）同时进行优化。

参数染色体采用 16 位二进制编码，C 和 σ^2 各占 8 位。每一个选中的参数都用 SVM 来评价。特征灵敏度可以用来反映特征参数对输出变化的敏感度。这里将特征集 x 对输出 y 的灵敏度定义为

$$\varepsilon(x \mid y)=|\nabla f(x)|=\left|\sum \alpha_i y_i \nabla_x K(x_i,x)\right| \tag{5}$$

其中

$$\nabla_x K(x_i,x)=(x_i-x)\sigma^2\exp(-0.5\cdot\|x_i-x\|^2\sigma^{-2}) \tag{6}$$

在这里，特征子集染色体采用二进制矢量表示，如 $V_{li}=\{v_1,v_2,\cdots,v_n\}$，$i=1,2,\cdots,n$ 表示第 l 类的第 i 位，若 $V_{li}=1$ 表示被选中，若 $V_{li}=0$ 表示未被选中。为了防止算法过早收敛，在 GA 中加入特征子集维数惩罚函数 $\zeta|x|$（其中 $|x|$ 表示特征子集的维数；n 表示特征子集总维数，$0<\zeta<1$[6]）。因此，适应度函数为

$$f(x)=\varepsilon(x)+\frac{1}{n}\zeta|x| \tag{7}$$

它的算法流程图如图 2 所示。

4 仿真实验

本文实验部分分为 2 块：第一部分主要是用 UCI 数据库中数据对算法性能进行样本训练；第二部分主要运用训练好的学习机进行样本测试和数据分析。

为了验证本文提出的算法通用性，特从 UCI 数据库上选取多类别数据，一共 1 728 个，每个样本有 6 个属性，4 个种类，去除 28 个含有异常属性样本后，还剩下 1 700 个，分为 2 类：一类为 Training，共 800 个，另一类为 Testing，共 900 个。本仿真实验采用基于 Matlab 开放源码 SVM[7]，它支持多分类问题。

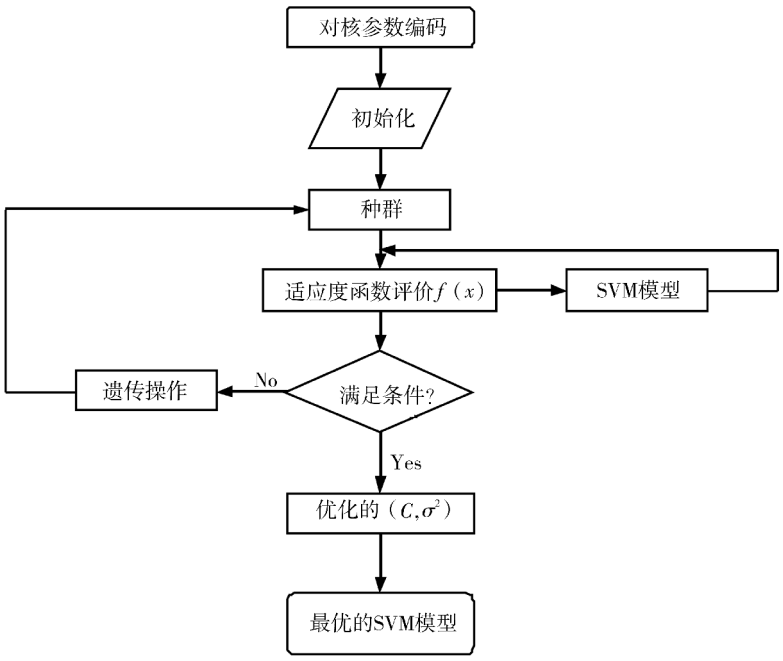

图 2　GA 优化的 SVM 模型

4.1 机器学习（样本训练）

即通过对样本数据的检测，求出各个分类函数的过程，算法如下：

Step1　根据 UCI 数据集每类占总样本数概率大小列出分类编号，见表 1。

Step2　对所测数据运用第 1 节提出 GA 进行预处理，去掉那些相关性强或者不易判别的特征。

Step3　构建基于二叉树结构的 SVM 多分类器，并给出分类依据。

Step4　运用第 3 节所给的方法，对分叉点的每个分类器两个参数（C、σ^2）进行优化，选出最佳结果。

Step5　从较多的训练样本中选择最优支持向量机[5]。

Step6　记录每一步实验的相关参数，若测试效果不理想的话，回到 Step1 调整记录的参数直至得到理想结果。

4.2 机器测试（样本测试）

使用基于二叉树结构双优化的 SVM 多分类器对实例进行分类时，由根节点开始对该对象的特征参数代入分类函数逐渐测试其值，当测试值为 1 时，即到达叶结点（终止前进，表示该分类工作正处于叶结点所代表的类

别);当测试值为-1时,则顺着分支向下走,直至到达某个叶结点。

表1 根据UCI原数据集先验概率编制表

识别类型	UNACC	ACC	GOOD	V-GOOD
数量/个	1 210	384	69	65
所占比例/%	70.023	22.222	3.993	3.762
编号	1	2	3	4

设置遗传算法的初始种群数为100,杂交率为0.90,变异概率0.06,迭代次数为80,惩罚系数$\zeta=0.16$[8-9]。计算最优适应度为0.523,在灵敏度值寻优过程中达到最优适应度时,灵敏度值保持不变。

表2 不同分类方法的实验精度和所用时间

分类方法	所用时间/s	分类精度/%	拒识别率/%
一对一法	274.85	91.86	4.94
有向无环图法	283.13	92.90	1.61
双优化二叉树法	268.65	95.78	2.98

从表2可以看出,在用相同样本数据作为训练和测试的情况下,本文提出的基于二叉树结构双优化的SVM多分类算法,不仅在测时间上少于上面两种方法,而且在分类精度上也比较高;然而在拒绝识别率上,本文提出的算法高于有向无环图法,但是远低于一对一法。这主要是由于笔者在实验的过程中,为了提高分类精度而选择适应度值偏大而造成的。通过上面的实验数据和分析来看,能够很好地说明本文提出的方法的可行性。

5 结论

本文提出了一种基于二叉树结构双优化的SVM多分类算法,不仅做到了对特征参数的优化,而且做到了对SVM分类器核参数的优化,从而大大降低了获取特征参数选择和核参数的时间[10-11]。在对UCI数据库中数据做的实验,达到了预期的效果,表明该方法的正确性和有效性。本文提出的这种算法在工程应用中,可以为设备故障智能诊断提供一种很好的新方法,这也将是笔者将来拓展的领域之一。

参考文献:

[1] 马笑潇,黄席樾,柴毅.基于SVM的二叉树多分类算法及其在故障诊断中的应用[J].控制与决策,2003,18(3):272-276.
Ma X X, Huang X Y, Cai Y. 2PTMC classification algorithm based on support vectormachines and its application to fault diagnosis[J]. Control and Decision, 2003, 18 (3): 272-276.

[2] 吴德,刘三阳.支持向量域多分类器[J].西安交通大学学报,2012,46(6):87-91.
Wu D, Liu S Y. Multiple support vector domain classifier [J]. Journal of Xi'an Jiaotong University, 2012, 46(06): 87-91.

[3] 袁胜发,褚福磊.支持向量机及其在机械故障诊断中的应用[J].振动与冲击,2007,26(11):29-35.
Yuan S F, Zu F L. Support vector machine and its application in machine fault diagnosis[J]. Journal of Vibration and Shock, 2007, 26(11): 29-35.

[4] 蒋维,钟小强,陈开,等.基于优化的支持向量机的机械设备多故障诊断模型[J].计算机应用与软件,2009,26(1):62-63.
Jiang W, Zhong X Q, Chen K, et al. Machine mault diagnosis of optimized support vector machine[J]. Computer Application and Software, 2009, 26(1): 62-63.

[5] 王宏勇,侯慧芳,刘素华.基于遗传算法和支持向量机的玉米品种识别[J].计算机工程与应用,2008,44(18):221-223.
Wang H Y, Hou H F, Liu S H. Maize seed recognition based on genetic algorithm and multi-class SVM[J]. Computer Engineering and Applications, 2008, 44 (18): 221-223.

[6] 王新峰,邱静,何正嘉.基于支持向量机的多故障分类器及其应用[J].机械工程学报,2006,42(4):122-126.
Wang X F, Qiu J, He Z J. Based on support vector machine fault classifier and its application[J]. Chinese Journal of Mechanical Engineering, 2006, 42(4): 122-126.

[7] 胡良谋,曹克强,徐浩军,等.支持向量机故障诊断及其控制技术[M].北京:国防工业出版社,2011.
Hu L M, Cao K Q, Xu H J, et al. Support vector machine fault diagnosis and control technology[M]. Beijing: National Defense Industry Press, 2011.

[8] 李学洋,李悦,张亚伟.基于遗传变异蚁群算法的机器人路径规划的改进[J].电子设计工程,2012,20(15):38-43.

Li X Y, Li Y, Zhang Y W. Improved ant colony algorithm based on genetic variation apply in robots path planning [J]. Electronic Design Engineering, 2012, 20(15): 38-43.

[9] 鲍雄伟. 小波变换在图像边缘检测中的应用[J]. 电子设计工程, 2012, 20(14): 160-162.
Bao X W. The application of wavelet transform in image edge detection[J]. Electronic Design Engineering, 2012, 20(14): 160-162.

[10] 徐国浪, 魏延. 基于多核函数的模糊支持向量机学习算法[J]. 重庆师范大学学报: 自然科学版, 2012, 29(6): 50-53.
Xu G L, Wei Y. Learning algorithm based on fuzzy support vector machine of multi-core functions[J]. Journal of Chongqing Normal University: Natural Science, 2012, 29(6): 50-53.

[11] 邬啸, 魏延, 吴霞. 改进的双隶属度模糊支持向量机[J]. 重庆师范大学学报: 自然科版, 2011, 28(5): 49-52.
Wu X, Wei Y, Wu X. Improved double memberships of Fuzzy support vector machine[J]. Journal of Chongqing Normal University: Natural Science, 2011, 28(5): 49-52.

Based on the Binary Tree Structure Double Optimization SVM Classification Algorithm

XU Guo-lang[1], WEI Yan[2]

(1. College of Mathematics;
2. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: Because of classification accuracy of the traditional binary tree for multi-classification problems is not high and it is too high for the time complexity, the authors of this paper present a new double optimization learning algorithm, based on the binary tree structure, which is a multi-classification algorithm. It makes the best of genetic algorithm to make feature parameters subset and kernel parameters optimized, in order to acquire the best important characteristic parameter combination for the purpose, and it can effectively solve the program of identification of complicated structure and uneven distribution sample. Combining with the UCI data in a database, through the simulation experiment, and compare the accuracy and time complexity with directed un-acyclic graph and one-to-one method, and the results show that the algorithm which has been proposed by the authors is effective in this paper.

Key words: GA; SVM; binary tree; multi-classification identification

(责任编辑 游中胜)