

文章编号:1000-5013(2016)02-0171-04

doi:10.11830/ISSN.1000-5013.2016.02.0171

采用粒子群优化的 SVM 算法在 数据分类中的应用

邹心遥¹, 陈敬伟¹, 姚若河²

(1. 广东农工商职业技术学院 机电系, 广东 广州 510507;

2. 华南理工大学 电子与信息学院, 广东 广州 510641)

摘要: 针对分类数据集合线性不可分的问题,改进了支持向量机(SVM)的分类方法,构建新的分类决策函数和高斯核函数.在支持向量机关键参数的优化环节,采用粒子群算法对惩罚参数和高斯参数进行优化,设计便于操作的优化流程,并针对 Iris 数据集展开实验研究.结果表明:相比于基于遗传算法优化的 SVM 方法,所提出的方法执行速度快、分类准确率高.

关键词: 数据分类; 支持向量机; 粒子群优化; Iris 数据集; 惩罚参数; 高斯参数

中图分类号: TP 181

文献标志码: A

数据分类作为数据库管理系统的核心技术,已经成为国内外学者广泛关注的焦点研究^[1].通过数据分类,可以将数据库中的各项数据分门别类,分类后的数据将具有明显的内在联系或意义近似性,更利于数据管理与维护,尤其是缩小数据查找的范围^[2].已经成功用于数据分类的方法很多,如邻近算法、贝叶斯算法、决策树算法、神经网络算法等^[3-5].Agarwal 等^[6]在 Fisher 核函数的基础上进行改进,形成一种含有概率特征的离散化形式,对数字数据挖掘具有较强的针对性.Durduran^[7]构建一个新的核函数空间,并设计在全空间上进行搜索的分类策略.李婷等^[8]以车载激光点云数据为分类对象,借助混合核函数的理念,对传统的高斯核函数进行修正,建立一种新的支持向量机(SVM)分类方法.陆慧娟等^[9]面向基因表达数据,针对局部分类采用径向基函数(RBF)核函数,针对全局分类采用线性核函数.支持向量机方法的关键在于如何对关键参数优化,如果能用智能算法实现这些参数的优化,并获得最优结果,就可以获得更高性能的 SVM 分类结果^[10].本文选择支持向量机的方法作为数据分类的研究对象,并在参数优化环节中选取粒子群(PSO)算法,构建出一种新的数据分类方法.

1 支持向量机分类方法

在线性可分的情况下,支持向量机分类方法通过在高维空间构造最优分类超平面,以最低的结构风险进行分类,具有学习能力强、训练速度快、分类精度高等诸多优点.然而,很多分类问题中的数据集合并不是线性可分的.此时,数据分类过程中的优化问题需要的数学模型定义为

$$\left. \begin{aligned} \min_{z, v} \quad & \frac{1}{2} v^T v + \rho \sum_{i=1}^n \eta_i, \\ \text{s. t.} \quad & q_i(v^T \cdot p_i + \rho) \geq 1 - \eta_i. \end{aligned} \right\} \quad (1)$$

式(1)中: (p_i, q_i) 为第 i 个训练样本, $i=1, 2, \dots, n$; $v \in \mathbf{R}^d$ 为 SVM 分类中超平面的法向向量; $z \in \mathbf{R}$ 为阈值; ρ 为惩罚参数,其值越大表示对不正确分类结果的惩罚程度越大,但过大的 ρ 值会降低 SVM 方

收稿日期: 2015-12-22

通信作者: 邹心遥(1978-),女,副教授,博士,主要从事新型光电器件、物联网技术的研究. E-mail: madelinexy@163.com.

基金项目: 国家自然科学基金资助项目(61274085); 广东省大学生科技创新培育专项(PDJH2015A0718)

法的泛化能力,所以 ρ 的选取要在分类精度和泛化能力之间寻求平衡.

式(1)是一个典型的二次规划问题,对此问题的求解可以采用拉格朗日乘子算法,即

$$L = \frac{1}{2} \mathbf{v}^T \mathbf{v} + \rho \sum_{i=1}^n \eta_i - \sum_{i=1}^n \theta_i [q_i (\mathbf{v}^T \cdot \mathbf{p}_i + z) - 1 + \eta_i] - \sum_{i=1}^n \vartheta_i \eta_i. \tag{2}$$

式(2)中: θ_i, ϑ_i 都是不小于 0 的拉格朗日乘子.

针对式(2),分别对变量 \mathbf{v}, z, η_i 求偏导,并让 3 个偏导值为 0,将此时求得的结果代入式(1),可得

$$\left. \begin{aligned} \min_{\theta} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n q_i q_j \theta_i \theta_j (\mathbf{p}_i \cdot \mathbf{p}_j) - \sum_{i=1}^n \theta_i, \\ \text{s. t.} \quad & \sum_{i=1}^n \theta_i q_i = 0, \quad 0 \leq \theta_i \leq \rho. \end{aligned} \right\} \tag{3}$$

对于式(3),能够满足 $\theta_i \leq \rho$ 的情况称为支持向量. 据此,最终可以得到用于分类的决策函数为

$$F(\mathbf{p}) = \text{sgn}(\sum_{i=1}^n q_i \theta^* (\mathbf{p}_i \cdot \mathbf{p}) + z^*). \tag{4}$$

要判断任意一个样本数据的类别情况,只要将其代入式(4)进行计算即可.

核函数是构造 SVM 分类方法的另一个关键问题. 因为考察的是一般分类情况,所以选取高斯核函数作为文中 SVM 分类方法的核函数,其简化形式为

$$H(\mathbf{p}_i, \mathbf{p}_j) = \exp[-\kappa \|\mathbf{p}_i - \mathbf{p}_j\|^2]. \tag{5}$$

式(5)中:参数 $\kappa=1/\sigma^2$, σ 称为高斯核函数的高斯系数.

2 SVM 分类中的粒子群参数优化

要提升 SVM 分类方法的精度和效率,对关键参数的设置和优化至关重要^[11]. 因此,采用粒子群算法对 2 个关键参数进行优化调整,以达到最佳的分类效果^[12]. 粒子群算法在执行过程上类似于遗传算法,也需要对粒子群进行初始化,并根据适应度函数进行优化操作. 相比于遗传算法,粒子群算法的实现更为简单,一方面,粒子群算法无需执行交叉和变异操作,另一方面,粒子群算法需要调整的参数没有遗传算法多. 基于这些情况,粒子群算法可以以更快的速度达到收敛、获得最优解^[13-15].

粒子群算法的基本实现过程如下.

在一个维度为 m 的待查找空间上,假设粒子群的初始状态为 $S=(s_1, s_2, \cdots, s_n)$, s_i 代表优化问题的第 i 个解,它在待查找空间上的多维表示为 $s_i=(s_{i,1}, s_{i,2}, \cdots, s_{i,m})$. 若各个粒子的速度为 $\mathbf{V}_i=(v_1, v_2, \cdots, v_m)^T$,群内的粒子个体位置最优解为 $\mathbf{L}_i=(l_{i,1}, l_{i,2}, \cdots, l_{i,m})^T$,能使种群整体达到最优状态的解为 $\mathbf{L}_g=(l_{g,1}, l_{g,2}, \cdots, l_{g,m})^T$,找到当前状态下的 \mathbf{L}_i 和 \mathbf{L}_g 以后,需要对群内的粒子进行位置更新和速度更新,其处理过程为

$$s_{i,d}(t+1) = s_{i,d}(t) + v_{i,d}(t+1), \tag{6}$$

$$v_{i,d}(t+1) = G \cdot v_{i,d}(t) + a_1 r_1 \cdot (\mathbf{L}_i(t) - s_{i,d}(t)) + a_2 r_2 \cdot (\mathbf{L}_g(t) - s_{i,d}(t)). \tag{7}$$

式(6), (7)中: G 为粒子飞行过程中的惯性权重; a_1, a_2 为加速度参数; r_1, r_2 为随机数.

对于文中 SVM 分类方法,要优化的参数主要有惩罚参数 ρ 和 $\kappa=1/\sigma^2$. 根据粒子群优化方法的执行过程,为这 2 个参数的优化设置了 6 个优化步骤.

步骤 1 用 (σ, κ) 构建每一个粒子的初始状态,并完成种群内全部 n 个粒子的位置初始化和速度初始化,同时设定 2 个加速度参数 a_1, a_2 的数值为 2,以及最大迭代次数 T .

步骤 2 根据当前的各个粒子状态,计算适应度,并以 SVM 分类正确的比率作为评价各个粒子适应度高低的依据.

步骤 3 比较每一个粒子的适应度和其出现过的最优状态,选出 \mathbf{L}_i 并更新最优状态,以便于下一次迭代过程的比较.

步骤 4 比较各个粒子的适应度和整个种群出现过的最优状态,选出 \mathbf{L}_g 并更新最优状态,以便于下一次迭代过程的比较.

步骤 5 根据式(6), (7)更新粒子的位置和速度.

步骤 6 不断重复从步骤 2 到步骤 5 的工作,直到迭代过程达到最大迭代次数 T ,此时的 \mathbf{L}_g 下各个粒子的状态,就是最终优化的结果.

3 实验结果与分析

在分类方法的验证性实验中,选取数据分类领域中常用的标准测试数据集 Iris. 该数据集是根据鸢尾花这种植物的外形特征构建数据集,包含 3 个种类的鸢尾花,分别是山鸢尾花(*Iris setosa*)、杂色鸢尾花(*Iris versicolour*)、维吉尼亚鸢尾花(*Iris virginica*). Iris 数据集从鸢尾花的花萼长度特征、花萼宽度特征、花瓣长度特征和花瓣宽度特征等 4 个方面对上述 3 种鸢尾花进行特征区分. 整个数据集内共含有 50 个样本数据,有标准的正确分类作为分类方法分类结果测试的判断依据.

在分类性能测试过程中,采用了常见的交叉验证手段. 即将 Iris 数据集随机分作 J 个子集合,先用 1 到 $(J-1)$ 和子集对分类方法进行训练,确定分类决策模型后,对第 J 个子集进行分类以考察分类模型的正确性;然后,用第 2 到 J 个子集对分类方法进行训练,确定分类决策模型后,对第 1 个子集进行分类以考察分类模型的正确性;以此类推,遍历所有可能的交叉验证.

在分类方法的参数设置上,设定初始种群的规模是 20 个粒子,设定加速度参数 a_1, a_2 为 2,粒子飞行过程中的惯性权重 G 为 1,设定最大迭代次数 T 为 100. 对于交叉子集 J 的设定则分别设置为 3, 5, 7, 9 这 4 种情况. 同时,选择遗传算法优化的 SVM 分类方法(简称 SVM-GA),作为提出的粒子群算法优化的 SVM 分类方法(简称 SVM-PSO)的对比方法. 两种方法的分类准确率和执行时间的对比结果,如表 1 所示. 表 1 中: η 为分类准确率; t 为执行时间.

由表 1 可知:所提出的基于粒子群优化的 SVM 分类方法的分类准确率明显优于基于遗传算法的 SVM 分类方法. 随着交叉验证的分类子集数目增多,文中方法的分类准确率一直保持在 90% 以上. 由表 1 还可知:所提出的基于粒子群优化的 SVM 分类方法的执行时间明显低于基于遗传算法的 SVM 分类方法;随着交叉验证的分类子集数目增多,基于遗传算法的 SVM 分类方法的执行时间增加较为明显,而文中方法的执行时间增加幅度则不大.

表 1 不同分类方法准确率和执行时间对比数据

分类子集	$\eta/\%$		t/s	
	SVM-GA	SVM-PSO	SVM-GA	SVM-PSO
$J=3$	91.2	98.8	0.587	0.097
$J=5$	88.7	96.7	0.802	0.201
$J=7$	83.4	92.4	2.355	0.327
$J=9$	82.6	91.9	5.624	0.415

4 结束语

在考察数据集线性不可分的情况下,设计 SVM 方法的分类决策函数和高斯核函数. 为了进一步提升 SVM 方法的分类性能,采用粒子群算法对 SVM 方法的 2 个重要参数进行优化,包括对惩罚参数的优化和高斯系数的优化. 以 Iris 为分类实验的测试数据集,并选择基于遗传算法优化的 SVM 方法作为比较方法. 数据分类的验证实验表明,所提出的基于粒子群算法优化的 SVM 分类方法,具有更高的分类准确性和更快的分类速度.

参考文献:

[1] PALACIOS A, MARTINEZ A, SANCHEZ L. Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data[J]. Engineering Applications of Artificial Intelligence, 2015(44):10-24.

[2] 范士俊, 张爱武, 胡少兴, 等. 基于随机森林的机载激光全波形点云数据分类方法[J]. 中国激光, 2013, 40(9):1-7.

[3] 刘红岩, 陈剑, 陈国青. 数据挖掘中的数据分类算法综述[J]. 清华大学学报(自然科学版), 2002, 42(6):727-730.

[4] 杨帆, 林琛, 周奇凤, 等. 基于随机森林的潜在 k 邻近算法及其在基因表达数据分类中的应用[J]. 系统工程理论与

实践,2012,32(4):815-825.

[5] TABKHI S,NAJAFI A,RANJBAR R. Gene selection for microarray data classification using a novel ant colony optimization[J]. Neurocomputing,2015,168:1024-1036.

[6] AGARWAL S K,SHAH S,KUMAR R. Classification of mental tasks from EEG data using backtracking search optimization based neural classifier[J]. Neurocomputing,2015,166:397-403.

[7] DURDURAN S S. Automatic classification of high resolution land cover using a new data weighting procedure: The combination of K-means clustering algorithm and central tendency measures (KMC-CTM)[J]. Applied Soft Computing,2015,35:136-150.

[8] 李婷,詹庆明,喻亮. 基于地物特征提取的车载激光点云数据分类方法[J]. 国土资源遥感,2012,92(1):17-21.

[9] 陆慧娟,安春霖,马小平,等. 基于输出不一致测度的极限学习机集成的基因表达数据分类[J]. 计算机学报,2013,36(2):341-348.

[10] ELYASIGOMARI V,MIRJAFARI M S,SCREEN H R C. Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization[J]. Applied Soft Computing,2015,35:43-51.

[11] 喻小光,陈维斌,陈荣鑫. 一种数据规约的近似挖掘方法的实现[J]. 华侨大学学报(自然科学版),2008,29(3):370-374.

[12] 王江海,武林仙,吴扬扬. 基于刻画得数据空间数据源管理子系统[J]. 华侨大学学报(自然科学版),2012,33(5):509-513.

[13] 彭京,唐常杰,元昌安,等. 一种基于概念相似度的数据分类方法[J]. 软件学报,2007,18(2):311-322.

[14] 杨帆,林琛,周绮凤. 基于随机森林的潜在 k 近邻算法及其在基因表达数据分类中的应用[J]. 系统工程理论与实践,2012,32(4):815-826.

[15] 陆慧娟,安春霖,马小平,等. 基于输出不一致测度的极限学习机集成的基因表达数据分类[J]. 计算机学报,2013,36(2):341-348.

Application of SVM Algorithm Based on Particle Swarm Optimization in Data Classification

ZOU Xinyao¹, CHEN Jingwei¹, YAO Ruohe²

(1. Department of Mechanical and Electronic, Guangdong AIB Polytechnic College, Guangzhou 510507, China;

2. School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: According to the problem that the classification data set can not be divided, the classification method of support vector machine (SVM) is improved, and the new classification decision function and Gauss kernel function are constructed. Using particle swarm algorithm to optimize the penalty parameters and Gauss parameters, the optimization process is easy to operate. To experimental study the Iris data set, the results show that compared with the SVM method based on genetic algorithm, the proposed method performs fast and has high classification accuracy.

Keywords: data classification; support vector machine; particle swarm optimization; Iris data set; penalty parameter; Gauss parameter

(责任编辑: 钱筠 英文审校: 吴逢铁)