

# 基于蚁群算法的SVM模型选择研究

倪丽萍,倪志伟,李锋刚,潘永刚

(合肥工业大学 管理学院,安徽 合肥 230009)

**摘 要:**为了提高SVM的分类性能,提出使用蚁群算法来指导SVM模型参数的选择,并针对采用RBF作为核函数的SVM进行了实验。然后将该方法与基于GA的SVM模型选择方法进行了比较。实验证明采用蚁群算法具有一定的优势,它能在较短的时间内寻找到最优解,且最终得到的分类结果优于遗传算法。

**关键词:**支持向量机;模型选择;蚁群算法

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2007)09-0095-04

## SVM Model Selection Based on Ant Colony Algorithm

NI Li-ping, NI Zhi-wei, LI Feng-gang, PAN Yong-gang

(School of Management, Hefei University of Technology, Hefei 230009, China)

**Abstract:** In order to improve the performance of classifiers of SVM, this paper introduces ant colony algorithm to guide the selection of SVM model parameters with RBF kernel. This method is compared with SVM model selection method based on GA method. The experiment result shows ant colony algorithm can get the optimization solution in shorter time and higher classification accuracy than GA.

**Key words:** SVM; model selection; ant colony algorithm

### 0 引言

SVM是20世纪90年代中期由Vapnik和他的At & T Bell实验小组提出的,目的在于克服传统机器学习方法中存在的推广能力不强的现象。SVM以结构风险最小化为理论原则,并通过引入内积的概念,把复杂的非线性问题转化为线性问题,特别适用于解决小样本问题。SVM自提出以来,得到了很大发展,目前在模式识别、回归估计等方面都得到了广泛的应用。

在采用SVM进行求解问题时,模型的选择至关重要。大量的实验证明不同的核函数参数以及支持向量机中的二次规划参数都会影响到最终的分类准确率。目前对SVM模型选择的方法有以下几种:

(1)实验法。即尝试不同的参数对进行测试、分析比较后得出一个最适合问题的参数对。这样的选择方法缺乏指导,具有一定的随机性。

(2)网格法。它的主要思想是在设定的网格密度范围内遍历每一个点,将每一个点作为分类器的参数

进行测试,最后选择准确率最高的参数对<sup>[1]</sup>。这种方法耗时,当数据集很大时该方法显然不可操作。

(3)数值方法,例如,拟牛顿法、梯度下降算法等。这些方法对于初始值的选择很敏感。

(4)遗传算法。该方法是近两年才提出的,实验证明利用遗传算法的特点,来进行模型参数的自动选择能够在一定程度上提高分类的准确率<sup>[1-4]</sup>,但是遗传算法往往容易陷入局部极小值点,且运行周期较长。

蚁群算法是一种集体智能算法,相对于其他的智能优化算法而言具有正反馈性能,能够加速优化进程。因此为了进一步提高运行效率和分类准确率,笔者提出使用蚁群算法来指导参数的选择。

### 1 SVM简介

#### 1.1 线性支持向量机

对于一组给定的训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ,其中 $y_i \in \{-1, 1\}$ ,在线性可分的情况下,分类的目的在于寻求参数对 $(\omega, b)$ ,构造超平面 $(\omega \cdot x) + b = 0$ ,使得该超平面能将训练样本中的正类输入和负类输入很好地区分开。最终最优分类面的求解可以归结为下列最优化问题:

$$\min \frac{1}{2} \| \omega \|^2$$

收稿日期:2006-12-18

**基金项目:**安徽省自然科学基金资助项目(050460402);安徽省教育厅课题资助项目(2006sk010)

**作者简介:**倪丽萍(1981-),女,安徽合肥人,博士研究生,研究方向为机器学习、数据挖掘;倪志伟,博士,教授,博导,研究方向为机器学习、数据挖掘、人工智能。

$$\text{s. t. } y_i[(w \cdot x_i) + b] \geq 1 \quad i = 1, \dots, l \quad (1)$$

即最佳分类面应使两类样本到该平面的距离最大。由于考虑到可能存在有噪声的数据,分类面不能将这些点很好地分开,故引入松弛变量  $\xi_i \geq 0, i = 1, \dots, l$  和惩罚因子  $C$ , 其中  $\xi_i$  描述了训练集被错划的程度,  $C$  控制的是训练误差和模型复杂度的折衷<sup>[5]</sup>。从而(1)式转换成:

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t. } & y_i[(w \cdot x_i) + b] \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \dots, l \end{aligned} \quad (2)$$

根据最优化理论,可以将(2)式转化为 Wolfe 对偶问题来求解:

$$\begin{aligned} \max w(a) = & \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l a_i a_j y_i y_j (x_i \cdot x_j) \\ \text{s. t. } & 0 \leq a_i \leq C, i = 1, \dots, l \end{aligned} \quad (3)$$

$$\sum_{i=1}^l y_i a_i = 0$$

求得最优解为  $w^* = \sum_{i=1}^l a_i^* x_i y_i$ , 其中  $a_i^*$  不为 0, 其所对应的样本称为支持向量。最终的判别函数为:

$$f(x) = \text{sgn}((w^* \cdot x) + b^*) = \text{sgn}(\sum_{i=1}^l a_i^* y_i (x_i \cdot x) + b^*)$$

## 1.2 非线性支持向量机

对于非线性可分的情况,通过引入核函数把非线性可分的输入空间映射到线性可分的空间。因而上述(3)式转化为:

$$\begin{aligned} \max w(a) = & \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l a_i a_j y_i y_j K(x_i, x_j) \\ \text{s. t. } & 0 \leq a_i \leq C, i = 1, \dots, l \\ & \sum_{i=1}^l y_i a_i = 0 \end{aligned}$$

同理可以求得判别函数为:

$$f(x) = \text{sgn}(\sum_{i=1}^l a_i^* y_i K(x_i, x) + b^*) \quad (4)$$

其中  $K(x_i, x)$  是核函数。目前核函数有多种类型,如线性核函数、多项式核函数、RBF 核函数等。文中以 RBF 核函数为研究对象, RBF 是径向基核函数。

$$k(x_i, x) = \exp(-\gamma \|x_i - x\|^2) \quad (5)$$

由(4)、(5)式可知基于 RBF 的 SVM 参数选择主要是惩罚因子  $C$  和核函数参数  $\gamma$ 。即选择  $C$  和  $\gamma$ , 使得由这两个参数确定的 SVM 分类器能够达到最好的分类效果。

## 2 基于蚁群算法的 SVM 模型选择

蚁群算法最早是由意大利学者 Dorigo 提出的,它

是一种集体智能算法。该算法虽然提出的时间较晚,但是发展很快,目前在求解 TSP、车间作业调度、网络路由(Qos)、背包问题等方面都得到了广泛应用,且仿真结果显示,蚁群算法具有较好的效果。

用蚁群算法进行 SVM 参数的选择是将蚁群算法用于解决连续域问题,即利用蚁群算法在连续空间中寻找最优参数  $C$  和  $\gamma$ , 使得所得到的分类器误差最小。算法流程如图 1 所示。

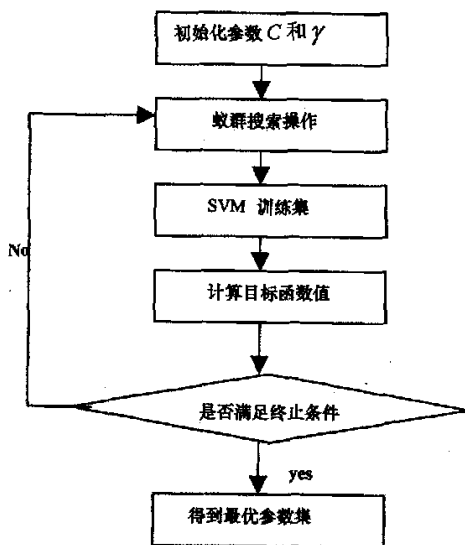


图 1 基于蚁群算法的 SVM 模型参数选择流程图

从该流程图中可以看出目标函数的选择和蚁群搜索操作的实现是基于蚁群算法 SVM 模型参数选择的两个关键问题。

### 2.1 目标函数的选择

SVM 模型参数的选择,目的在于最小化 SVM 的泛化误差。目前对 SVM 泛化误差的估计方法有多种,其中最常用的是交叉验证(cross-validation)和留一法(leave-one-out),但这两种方法效率都不高,尤其是在样本数量很多的情况下。文中采用 Joachims 方法<sup>[6]</sup>,该方法只需要一次训练,直接利用 SVM 的解,因而效率较高。

该方法的具体表示如下:

$$\begin{aligned} f &= E_n \hat{\xi} = \frac{d}{n} \\ d &= |\{i: (\rho \alpha_i R_\Delta^2 + \xi_i \geq 1)\}| \end{aligned}$$

其中  $d$  是用留一法所产生的误差个数上界;  $n$  是训练样本数, 越小, 所得的 SVM 性能越好。因此在运用蚁群算法时, 令目标函数  $F(C, \gamma) = f$ , 寻参的过程就是最小化  $F$ 。

### 2.2 蚁群搜索操作

蚁群算法的提出起源于离散型的路径问题, 因而

在求解连续域问题上要进行变化。目前在连续域上如何运用蚁群算法日益受到重视,很多文献都提出或改进了用于求解连续域空间上的蚁群算法<sup>[7,8]</sup>。其中文献<sup>[7]</sup>提出了一种 MG-CACO 方法,文中的蚁群搜索操作就是基于此方法,并进行了一定的改变。具体操作如下:

Step1. 根据目标函数的优劣,找出当前最优位置的蚂蚁,即当前最优解,称这只蚂蚁为头蚁。

Step2. 除头蚁外,选择最差的一半蚂蚁进行成群募集操作,剩余的一半执行海量募集操作。其中成群募集操作的作用是让头蚁引领较差的个体向当前最优的位置移动,而海量募集操作则是蚂蚁根据各自信息素大小选择其移动的位置。这两部分操作的具体描述可以参见文献<sup>[7]</sup>。

Step3. 头蚁进行局部搜索,搜索策略如下:

$$x_{\text{temp}}(t) = x_{\text{leader}}(t) + \alpha * \text{step} \quad p < 0.5$$

$$x_{\text{temp}}(t) = x_{\text{leader}}(t) - \alpha * \text{step} \quad p \geq 0.5$$

其中  $p$  是  $(0,1)$  上的随机数,  $\alpha$  随着迭代次数的增加而减少。如果  $x_{\text{temp}}(t)$  优于  $x_{\text{leader}}(t)$ , 则用此解替换,同时头蚁的信息素也要进行相应的更新,更新的式子可以表示成:

$$\tau(x_{\text{leader}}(t)) = \tau(x_{\text{leader}}(t)) + k[f(x_{\text{leader}}(t)) - f(x_{\text{temp}}(t))]$$

这一步与 MG-CACO 方法不同。MG-CACO 方法对每只蚂蚁都进行局部搜索,这样做会导致多次训练 SVM,降低寻优速度,因此文中在局部搜索中只对头蚁进行局部搜索。

Step4. 所有蚂蚁完成一轮搜索后要执行信息素挥发操作,即  $\tau(x, t) = \rho * \tau(x, t)$ ,  $\rho$  为挥发因子。

### 2.3 基于蚁群算法的 SVM 模型参数选择算法描述

至此,基于蚁群算法的 SVM 参数选择算法可以描述如下:

Step1. 设置迭代的最大次数、终止条件、蚁群大小和蚁群算法中的一些初始化参数。

Step2. 初始化参数  $C$  和  $\gamma$ , 即随机产生初始蚂蚁位置,每一个蚂蚁的初始位置都确定了一对  $C$  和  $\gamma$  的值,并计算相应的目标函数值。

Step3. 如果满足终止条件则输出最优参数,否则执行第 4 和第 5 步操作。

Step4. 执行蚁群搜索操作。

Step5. 计算相应的目标函数值。

## 3 实验结果及分析

为了测试基于蚁群算法模型参数选择的有效性。

使用 Keerthi 提供的数据库,该数据库包含 5 个数据集<sup>[9]</sup>,且每个数据集都是两类分类问题。表 1 说明了这 5 个数据集的具体情况。

表 1 Keerthi 提供的 5 个数据集

数据集	训练集样本数	属性数	测试集样本数	类别数
banana	400	2	4900	2
splice	1000	60	2175	2
tree	700	18	11692	2
waveform	400	21	4600	2
image	1300	15	1010	2

### 3.1 参数设置

蚁群算法中参数设置如下:挥发因子  $\rho = 0.7$ ,  $\text{step} = 0.2$ ,成群募集中设  $p_{\text{GR}} = 0.5$ ,海量募集中设  $p_{\text{MR}} = 0.5$ ,信息素更新中比例因子  $k = 0.01$ 。同时文中对  $C$  和  $\gamma$  两个参数采用实数编码,其中  $C$  的搜索范围设置为  $[0,500]$ ,  $\gamma$  的搜索范围设置为  $[0.1,10]$ 。由于两者的搜索范围不同,在进行局部搜索时,邻域的初始设置也不同。 $C$  的邻域半径初始值为 10,而  $\gamma$  的初始值为 0.01,邻域半径随时间衰减。

蚂蚁种群大小取  $N = 10$ ,如果最大进化代数超过 50 代或连续 10 代最优解之差的绝对值小于 0.01,则终止寻参。

为了与基于遗传算法的 SVM 模型选择方法进行比较,对基于遗传算法的 SVM 模型选择也进行了测试。遗传算法中的参数设置如下:种群大小为 20,交叉概率为 0.8,变异概率为 0.05,寻参的终止条件与蚁群算法中的一样。

SVM 使用  $\text{svm}^{\text{libsvm}}$ <sup>[10]</sup>。整个实验环境为 Intel(R) P4 CPU 1.7GHz, 512MB RAM。为了更加准确地比较两种算法,对每个数据集均运行 10 次,然后比较结果。

### 3.2 结果分析

表 2 和表 3 分别是基于蚁群算法和遗传算法的 SVM 参数选择结果。

其中最优  $C, \gamma$  与最优分类准确率是相对应的,即最优分类准确率是该最优参数所对应的 SVM 模型在测试集上的分类准确率,平均运行时间是 10 次寻参的平均时间。从表 2 和表 3 的结果中可以看出,采用蚁

表 2 基于蚁群算法的 SVM 参数选择结果

数据集	最优 $C, \gamma$	最优分类准确率	平均运行时间
banana	(279.9451, 1.115)	87.47%	326s
splice	(349.194, 0.1001)	62.53%	262.5s
tree	(114.741, 0.2614)	87.34%	892.6s
waveform	(382.725, 0.1026)	88.91%	60.1s
image	(113.609, 0.1054)	97.82%	1164s

表 3 基于 GA 算法的 SVM 参数选择结果

数据集	最优 $C, \gamma$	最优分类准确率	平均运行时间
banana	(496.000, 1.0633)	87.14%	836.5s
splice	(138.609, 0.1001)	61.93%	461s
tree	(277.382, 0.2485)	86.95%	1056.1s
waveform	(117.500, 0.1)	88.91%	103.5s
image	(154.500, 0.1198)	97.82%	1423.3s

群算法进行 SVM 模型选择取得了较好的效果,它能在较短的时间寻找到最优参数,而且最终所得的分类准确率与采用 GA 算法所得到的最优分类准确率相当,在大部分情况下还有所提高。

图 2 是在 tree 数据集上,利用遗传算法和蚁群算法 10 次寻参后,所得分类器分类准确率的比较图。

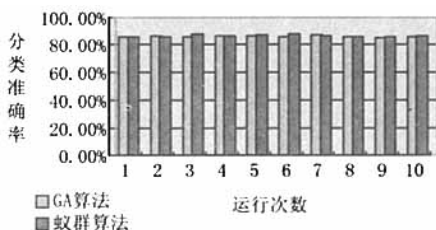


图 2 遗传算法和蚁群算法寻参比较图

从图中可以看出,10 次运行中利用蚁群算法所得的 SVM 分类器,最优分类准确率和平均分类准确率都要优于用遗传算法所得的结果。

除此之外,在实验当中还发现使用 Joachims 方法的  $E_{\gamma, \xi_a}$  作为目标函数估算错误率时较为准确,仍然以 tree 数据集为例,图 3 是在其上用蚁群算法寻找参数时的估算错误率与测试错误率(实际错误率)的曲线图。从图中可以看出两者的变化趋势基本一致。

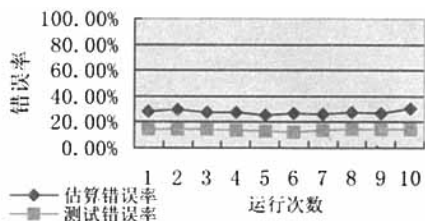


图 3 tree 数据集上预测误差与测试误差的曲线图

## 4 总 结

采用蚁群算法来进行 SVM 模型的参数选择。从实验的结果可以看出,蚁群算法较之遗传算法具有更好的效果,不论是分类准确率还是效率上都有一定的提高。下一步的工作是尝试采用其他的性能估计方法如:GACV, Radius - Margin bound 等作为目标函数,进行进一步的比较研究。

### 参考文献:

- [1] Chen Peng - Wei, Wang Jung - Ying, Lee Hahn - Ming. Model Selection of SVMs Using GA Approach[C]//Proceedings of the International Joint Conference on Neural Networks 2004. Hungary: IEEE, 2004: 2035 - 2040.
- [2] Xu P, Chan A K. Support vector machine for multi - class signal classification with unbalanced samples[C]//Proceedings of the International Joint Conference on Neural Networks 2003. Portland: IEEE, 2003: 1116 - 1119.
- [3] Xu P, Chan A K. An efficient algorithm on multi - class support vector machine model selection[C]//Proceedings of the International Joint Conference on Neural Networks 2003. Portland: IEEE, 2003: 3229 - 3232.
- [4] 黄景涛, 马龙华, 钱积新. 一种用于多分类问题的改进支持向量机[J]. 浙江大学学报: 工学版, 2004, 38(12): 1633 - 1636.
- [5] 董春曦, 饶 鲜, 杨绍全, 等. 支持向量机参数选择方法研究[J]. 系统工程与电子技术, 2004, 26(8): 1117 - 1120.
- [6] Joachims T. Estimating the generalization performance of an SVM efficiently[C]//Proceedings of the 17th International Conference on Machine Learning. San Francisco: Morgan, 2000: 431 - 438.
- [7] 贺益君, 俞欢军, 陈德钊. 基于募集机制的连续蚁群系统及其应用[J]. 浙江大学学报: 工学版, 2006, 40(5): 748 - 752.
- [8] 孙学勤, 刘 丽, 付 萍, 等. 一种连续空间优化问题的蚁群算法及应用[J]. 计算机工程与应用, 2005, 41(34): 217 - 220.
- [9] Keerthi S S. Benchmark datasets[EB/OL]. 2001 - 09 - 12. <http://guppy.mpe.nus.edu.sg/mpessk/comparison.shtml>.
- [10] Joachims T. SVMlight Support Vector Machine[EB/OL]. 2004 - 02 - 09. <http://svmlight.joachims.org>.

(上接第 94 页)

with Mobile Backbones[C]//Proceedings of IEEE International Symposium on Personal, Indoor and Radio Communications (PIMRC). Barcelona, Spain: IEEE Press, 2004: 566 - 573.

- [3] 王海涛. 移动 Ad hoc 网络的分簇算法及性能比较[J]. 北京邮电大学学报, 2004, 27(1): 93 - 97.
- [4] Chatterjee M, Das S K, Turgut D. A weighted clustering algo-

rithm (WCA) for Ad hoc networks[C]//Proceedings of IEEE GLOBECOM 2000. San Francisco: IEEE Press, 2000: 1697 - 1701.

- [5] Bettstetter C, Resta G, Santi P. The Node Distribution of the Random Waypoint Mobility Model for Wireless Ad Hoc Networks[J]. IEEE Transactions on Mobile Computing, 2003, 2(3): 257 - 269.