文章编号:1001 - 9081(2008)01 - 0134 - 02

利用粒子群算法优化 SVM 分类器的超参数

王 东[1,2],吴湘滨[2]
(1. 佛山科学技术学院 计算机科学与技术系,广东 佛山 528000; 2. 中南大学 地学与环境工程学院,长沙 410083)
(wdong@fosu.edu.cn)

摘 要:利用粒子群算法在求解组合优化问题时具有的全局搜索特性,设计并实现了支持向量机分类器中超参数的优选粒子群算法,扼要地叙述了算法实现中个体编码和适应度函数,通过在国际标准数据集上的实验验证了算法的有效性和高效性,最后列举了一些在上述工作基础上可开展的深入性工作。

关键词:支持向量机;分类器;参数优化;粒子群优化算法
中图分类号: TP18 **文献标志码**:A

Utilizing particle swarm optimization to optimize hyper-parameters of SVM classifier

WANG Dong[1,2], WU Xiang-bin[2]
(1. *Department of Computer Science and Technology, Foshan University, Foshan Guangdong* 528000, *China*;
2. *College of Geosciences and Environmental Engineering, Central South University, Changsha Hunan* 410083, *China*)

Abstract: Particle swarm optimization used for optimization selection for hyper-parameter of support vector machine classifier was designed and implemented utilizing global searching property of particle swarm optimization algorithm while the algorithm was used to solve combinatorial optimization problems. The method of individuals coding and evaluating was described in brief. The experimental statistic results demonstrate that the algorithm is effective and efficacious. In the end, some in-depth works are listed on the base of above-mentioned study.

Key words: Support Vector Machine (SVM); classifier; parameter optimization; particle swarm optimization

支持向量机(SVM)是在统计学习理论和结构风险最小基础上发展起来的一种分类方法,具有较强的泛化能力[1]。通过将非线性变换转换到高维特征空间,把待求解问题转化为二次优化问题,使 SVM 收敛于问题的全局最优解。

由于 SVM 分类器模型中参数的选取,对分类器的性能从精度和速度两方面产生较大的影响,同时为避免网格搜索超参数带来的时间消耗和搜索范围难于确定问题,许多研究提出利用智能算法对上述主要参数组合进行搜索,具有代表性的研究工作包括:文献[2]在分析参数对分类器识别精度的影响基础上,提出优化分类器参数的自适应遗传算法;文献[3]分析了 SVM 各参数对其性能的影响,根据已有的样本集确定遗传算法的搜索区间,然后在该区间内对搜索的参数进行最优选取;文献[4]利用免疫算法完成 SVM 参数和特征选择联合优化的方法;文献[5]利用蚁群算法进行 SVM 分类特征选取等。

本文以二类 SVM(*C*-SVM)为研究对象,通过对惩罚因子 C 和高斯核函数 RBF 中参数 σ 的基于网格的 k-折交叉验证值(以下简称验证值)实验结果分析,认为 *C*-SVM 参数优选问题为组合优化问题,由此提出并建立了搜索超参数粒子群算法,以提高 SVM 超参数搜索的效率。

1 支持向量机分类器

1.1 分类器模型

由于多类的分类问题可转化为二类的分类问题,为不失一般性,本文仅讨论二类问题。设训练向量 $\vec{x}_i \in R^n, i = 1,2,\cdots,l$ 属于二类,即 $y_i \in \{-1, +1\}$。带有松弛因子的分类目标可描述为以下初始问题:

$$\min_{w,b,\xi_i}\left(\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l}\xi_i^2\right) \tag{1}$$

$$\text{s.t.} \quad y_i(\langle\vec{w},\vec{x}_i\rangle + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1,2,\cdots,l$$

其中:$\vec{w}$ 为超平面 $[\vec{w},\vec{x}_i] + b = 0$ 的法矢量,b 为超平面偏值,C 为惩罚参数,ξ_i 为松弛因子表示 $\vec{x}_i$ 到超平面 $[\vec{w},\vec{x}_i] + b = 0$ 的距离。当训练集为非线性时,通过一个非线性映射 $\Phi(\vec{x}_i)$ 把训练数据 $\vec{x}_i$ 映射到一个高维线性特征空间。求解时并不需要计算该非线性函数,只需计算核函数 $K(\vec{x}_i,\vec{x}_j) = [\Phi(\vec{x}_i), \Phi(\vec{x}_j)]$。采用拉格朗日乘子法求解这个具有线性约束的二次规划问题,将原问题转化成对偶问题的最大值,则有:

$$\max_{\alpha}\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\vec{x}_i,\vec{x}_j) \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^{l}\alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C$$

采用不同的核函数 $K(\vec{x}_i,\vec{x}_j)$ 将导致不同的支持向量计算法。目前广泛应用的核函数形式主要有线性核函数、多项式核函数、高斯核函数(RBF)、Sigmoid 核函数。由于高斯核函数可逼近任意非线性函数,所以,本文以采用 RBF 的 SVM 进行研究。RBF 表示为:

收稿日期:2007 - 07 - 31;**修回日期**:2007 - 10 - 13。 **基金项目**:国家自然科学基金资助项目(40473029)。
作者简介:王东(1970 -),男,黑龙江甘南人,讲师,博士研究生,主要研究方向:地理信息系统、组合优化、智能计算; 吴湘滨(1962 -),男,湖南长沙人,教授,博士生导师,博士,主要研究方向:环境地质。

$$K(\vec{x}_i,\vec{x}_j) = \exp\left\{-\frac{|\vec{x}_i-\vec{x}_j|^2}{\sigma^2}\right\} \tag{3}$$

1.2 超参数选取对模型的影响

许多文献研究结果指出,在 C-SVM 中式(2)和(3)中惩罚因子 C 和 RBF 参数 σ 的选取,对分类性能将产生较大的影响。其原因在于,惩罚因子 C 用于控制模型复杂度和逼近误差,其值越大则对数据的拟合程度越高,但将导致泛化能力降低;σ 值(图 1 中为 γ)对模型的分类精度有重要影响。

下面给出采用 LIBSVM[6] 对上述两个参数在 3 个国际标准数据[7] 集上的统计结果,其中对每给定一组参数组合,计算其验证值并按照绘制验证值的等值线图,如图 1 所示。

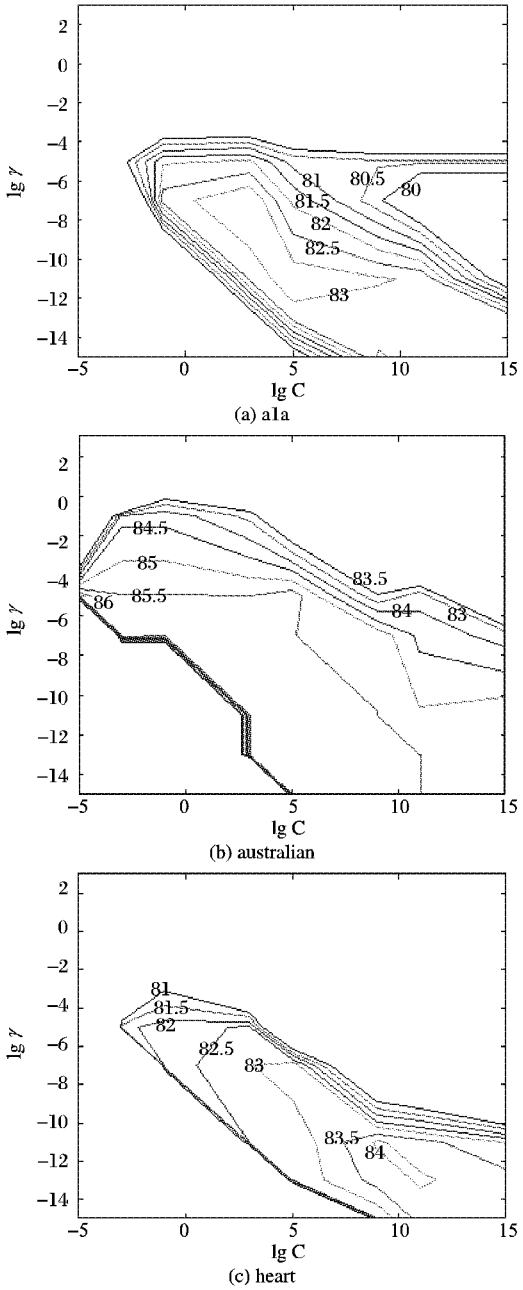


图 1　参数组合统计样例图

从图 1 可看出:

1)分类模型按照 C 和 σ 取值的组合情况不同,其验证值分布呈等值线且等值类之间无交叉现象。

2)根据验证值分布情况可知,对于上述两个参数组合优选,属于多峰值的组合优化问题,适宜于采用智能算法进行搜索。

2 粒子群优化算法

2.1 基本粒子群算法

设待搜索问题的解空间为 D 维空间,群中的粒子总数为 N。第 i 个粒子位置表示为向量 $\boldsymbol{X}_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(D)})$,该粒子已搜索到的最优位置为 $\boldsymbol{P}_i=(p_i^{(1)},p_i^{(2)},\cdots,p_i^{(D)})$,粒子群已搜索到的最优位置为 $\boldsymbol{P}_g=(p_g^{(1)},p_g^{(2)},\cdots,p_g^{(D)})$,第 i 个粒子的飞行变化率(速度)为 $\boldsymbol{V}_i=(v_i^{(1)},v_i^{(2)},\cdots,v_i^{(D)})$。每个粒子的速度和位置按如下公式飞行:

$$\begin{cases} v_i^{(d)}(t+1) = \omega\times v_i^{(d)}(t) + c_1\times\varphi_1\times[p_i^{(d)}(t) - x_i^{(d)}(t)] + c_2\times\varphi_2\times[p_g^{(d)}(t) - x_i^{(d)}(t)] \\ x_i^{(d)}(t+1) = x_i^{(d)}(t) + v_i^{(d)}(t+1) \end{cases} \tag{4}$$

其中:$1\leqslant i\leqslant N$,$1\leqslant d\leqslant D$;c_1 和 c_2 为大于 0 的常数,称为加速因子;φ_1 和 φ_2 为[0,1]之间的随机数;ω 称惯性因子,ω 较大适于对解空间进行大范围探查,ω 较小适于进行小范围开采。第 d 维的位置变化范围为 $[-XMAX_d,XMAX_d]$,速度变化范围为 $[-VMAX_d,VMAX_d]$,即在迭代过程中,若 $v_i^{(d)}$ 和 $x_i^{(d)}$ 超出了边界值,将之设为边界值。

2.2 优化超参数的 PSO

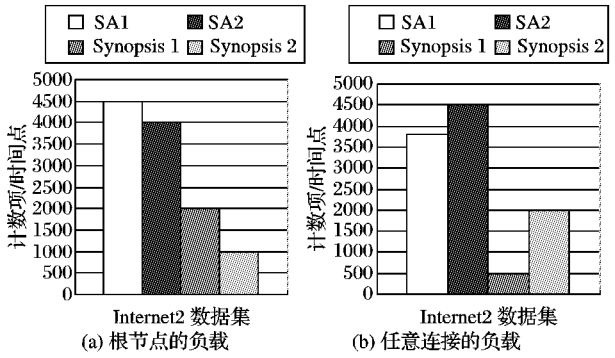
在利用 PSO 求解组合优化问题时有两个关键问题,一个是编码方式,一个是适应度评价方法。编码方式直接影响到算法搜索的效率和收敛速度,目前多数学者提出采用实数编码,个体则为二进制编码方式。通过对 LIBSVM 提供的网格搜索工具分析可知,其搜索方式采用的是整数编码方式,并且采用二重循环方式进行穷尽式搜索,当对取样的两个整数作为参数对 C-SVM 进行 k-折交叉验证时,取 2 的该整数次幂,即设当前搜索的 C 值为 c,则带入 C-SVM 时取 2^c。

有鉴于此,本文提出的参数优选算法采用整数编码方式。适应度函数在 PSO 算法求解问题过程中通过对个体的优劣评价引导算法搜索向问题的更优解收敛,通常情况下适应度评价函数针对不同的问题确立,对于 C-SVM 参数优选问题常用的评价方式是 k-折交叉验证方法。由此,优化 C-SVM 超参数的 PSO 算法中每个粒子为二维,分别对应 C 和 σ,算法描述如下:

算法:C-SVM 参数优选 PSO 算法
输入:样本数据集
输出:优选的 C 和 σ
Begin
1)　初始化
(1)　设定搜索参数的上下界和步长;
(2)　建立 C-SVM 参数集;
(3)　建立二维搜索的 PSO 种群并随机初始化每个粒子;
(4)　设定每个粒子的局部最优粒子为当前粒子;
(5)　循环对每个粒子 p 做
a)　$c\leftarrow 2^{p.C}$;$g\leftarrow 2^{p.G}$;
b)　用 c,g 带入 C-SVM 参数集计算每个粒子的 k-折交叉验证值作为其适应度;
(6)　设定当前种群中适应度最优个体为全局最优粒子;
2)　循环直到达到最大迭代次数或结束条件,做
(1)　根据式(4)中的第 1 个公式计算每个粒子的速度;
(2)　根据式(4)中的第 2 个公式调整每个粒子的位置;
(3)　根据对每个粒子 p 做
a)　$c\leftarrow 2^{p.C}$;$g\leftarrow 2^{p.G}$;

(下转第 139 页)

就很差。正如所期望的,Synopsis 1 策略在所有方法中性能最好。而无论是根节点或叶子节点的概要算法 Synopsis 1 和 Synopsis 2,其性能均比文献[2]和文献[3]的流频率项挖掘算法 SA1 和 SA2 有很大的提高。



(a) 根节点的负载　　(b) 任意连接的负载

图 3　算法比较

4　结语

本文研究了在分布式流数据中发现频繁项的策略和算法。其核心问题是如何更好地管理多个分布式节点所产生的局部概要结构的合并来获得好的近似性能。通过设置和优化在分层通信拓扑结构中精确梯度来实现比其他算法更好的时间敏感性,并能有效解决最差输入和非最差输入问题,较好地减少了通信负载,所研究的策略和算法可适用于不同数据分布的大规模分布式系统。

参考文献:

[1] BABCOCK B, OLSTON C. Distributed top-k monitoring[C]// Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2003:102 - 114.

[2] VITTER J S. Random sampling with a reservoir [J]. ACM Transactions on Mathematical Software, 1985, 11(1):37 - 57.

[3] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]// Proceedings of the Twentieth International Conference on Very Large Data Bases. Santiago: VLDB Press, 1994:77 - 89.

[4] JIN C, QIAN W, SHA C, *et al*. Dynamically maintaining frequent items over a data stream [C]// Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management. New Orleans: ACM Press, 2003:225 - 237.

[5] 王鹏,吴晓晨,王晨,等. CAPE——数据流上的基于频繁模式的分类算法[J]. 计算机研究与发展,2004,41(10):1677 - 1683.

[6] WHANG K Y, VANDER-ZANDEN B T, TAYLOR H M. A linear-time probabilistic counting algorithm for database applications [J]. ACM Transaction on Database Systems, 1990, 15(2):208 - 229.

(上接第 135 页)

b) 用 c,g 带入 C-SVM 参数集计算每个粒子的 k- 折交叉验证值作为其适应度;

c) 若 新位置的适应度优于局部最优粒子 则
用新粒子替换局部最优粒子;

(4) 若 种群中的最优粒子优于全局最优粒子 则
用种群中的最优粒子替换全局最优粒子;

3) 返回全局最优个体中的 C 和 σ;

End.

3　实验情况

下述实验环境为 Intel T2300E 1.66 GHz CPU,1 GB 内存,操作系统为 Windows XP。PSO 中基本参数设置为 $c_1 = 2$, $c_2 = 2$, $\omega = 1$, 粒子群数量为 20,最大迭代次数为 20。PSO 算法运行的结束条件是达到或超过网格算法搜索的 k-折交叉验证值即刻停止。下面对网格搜索算法和本文提出算法的运行情况对比,两个算法 SVM 参数设置除 C 和 σ 以外完全相同,两个参数的搜索范围均是从 -5 到 15,步长设置为 1。PSO 算法属于随机搜索算法,故该算法被重复 30 次,并取 30 次重复数据的最优值、最差值、平均值和平均运行时间,所得实验统计结果如表 1 所示。

表 1　网格搜索算法与 PSO 参数优选算法运行效果对比

数据集名称	网格搜索算法		本文算法				
	搜索时间	最优值	最好值	最差值	平均		
					搜索时间	搜索值	与网格搜索最优值的差值
ala	673	83.1153	84.1745	83.1153	139.658	83.4081	0.2591
australian	100	56.5217	57.6812	56.5217	28.410	56.7874	0.2657
fourclass	146	99.0179	99.6520	99.0179	21.339	99.3503	0.3324
heart	18	59.2593	59.6296	59.2593	7.815	59.2716	0.0123

从实验结果可以看出,采用本文提出的算法,均能在较网格搜索算法更短的时间内搜索到更优的参数组合,表明本文提出算法的有效性和高效性。即:

1)若设定相同的搜索时间,本文提出的算法能搜索到优于网格搜索算法的 C 和 σ 参数组合;

2)若设定目标 k- 折交叉验证值,本文提出的算法能以较网格搜索算法更短的时间,完成搜索。

4　结语

本文提出的 SVM 分类器参数优选粒子群算法,充分利用粒子群算法的全局搜索特性,实验表明该算法能以更高的效率实现 SVM 分类器参数的优选。今后的工作主要在以下方面展开:为提高随机搜索算法效率,建立相应的启发式策略,进一步提高算法的收敛速度;建立确定的搜索结束条件,使算法能在确定条件下结束;如何确定搜索空间,进一步扩大参数取值搜索空间,以利于算法能获得更优的参数组合。

参考文献:

[1] VAPNIK V. The nature of statistical learning[M]. New York: Springer-Verlag, 1995.

[2] 陈果. 基于遗传算法的支持向量机分类器模型参数优化[J]. 机械科学与技术, 2007, 26(3):347 - 350.

[3] 杜京义,侯媛彬. 基于遗传算法的支持向量回归机参数选取[J]. 系统工程与电子技术, 2006, 28(9):1430 - 1433.

[4] 周红刚, 杨春德. 基于免疫算法与支持向量机的异常检测方法[J]. 计算机应用, 2006, 26(9):2145 - 2147.

[5] 燕中, 袁春伟. 基于蚁群智能和支持向量机的人脸性别分类方法[J]. 电子与信息学报, 2004, 26(8):1177 - 1181.

[6] CHANG C C, LIN C J. LIBSVM - A library for support vector machines [CP/OL]. http://www.csie.ntu.edu.tw/~cjlin/, 2001.

[7] CHANG C C, LIN C J. LIBSVM data: Classification, regression, and multi-label [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, 2001.