支持向量机算法多目标模型选择

黄景涛,池小梅,马建伟
HUANG Jing-tao,CHI Xiao-mei,MA Jian-wei

河南科技大学 电子信息工程学院,河南 洛阳 471003
Electronic & Information Engineering College,Henan University of Science and Technology,Luoyang,Henan 471003,China
E-mail:hjt.haust@gmail.com

HUANG Jing-tao,CHI Xiao-mei,MA Jian-wei.Multi-object model selection of Support Vector Machine.Computer Engineering and Applications,2009,45(24):143-145.

Abstract: To meet the different focuses on the performances of Support Vector Machine(SVM) during application,the model selection of SVM is taken as a Multi-Object Optimization(MOO) problem.Particle Swarm Optimization(PSO) is applied to solve this MOO problem.The solution known as Pareto front is gained and one can select a concrete single solution from it according to own application needs,i.e.final model selection.The experiments on several datasets show that the method can gain Pareto front faster and the elements in the Pareto set can meet the need on generalization performance and training speed in SVM application.
Key words: Support Vector Machine(SVM);model selection;Multi-Object Optimization(MOO);Particle Swarm Optimization(PSO)

摘 要: 为适应支持向量机(Support Vector Machine,SVM)算法应用过程中的不同性能指标要求,将SVM算法的模型选择问题作为一个多目标优化(Multi-Object Optimization,MOO)问题进行处理。以改进的粒子群优化(Particle Swarm Optimization,PSO)算法对该多目标优化问题进行求解,得到其Pareto解集,在具体应用中根据实际需要从Pareto解集中选择适合的最优解作为支持向量机算法参数,实现支持向量机算法的模型选择。在几个数据集上的仿真实验表明,该方法能够较快地得到Pareto解集,解集中的参数组合能够满足对支持向量机算法速度和泛化能力的不同要求。
关键词: 支持向量机(SVM);模型选择;多目标优化(MOO);粒子群优化(PSO)
DOI: 10.3778/j.issn.1002-8331.2009.24.042 **文章编号:** 1002-8331(2009)24-0143-03 **文献标识码:** A **中图分类号:** TP391

1 引言

模型选择是许多基于数据统计算法在应用过程中不可回避的关键问题之一,在确定需要的模型类型之后,需要将其具体化,即确定模型的全部参数。算法的参数选择问题也即模型选择。对于大多数含有参数的算法来说,不同的算法参数具有不同的性能特征,适用于不同的场合,因此,如何确定特定应用背景下的具体算法模型就成为算法开发和应用过程中必须关注的主要问题之一。传统的模型选择准则和方法主要有AIC(Akaike Information Criteria)、BIC(Bayesian Information Criterion)、MDL(Minimum Description Length)、CV(Cross Validation)等[1]。支持向量机算法(Support Vector Machine,SVM)是Vapnik[2]提出的一种基于结构风险最小化准则的小样本学习机,建立在统计学习理论基础之上,能够获得较好的泛化能力。核变换的引入,使其能够处理复杂的非线性问题,近年来在众多领域得到了应用。在SVM及其改进算法中,都涉及到算法参数的选择问题,即模型选择问题。

SVM的参数选择问题已经有不少学者进行了研究。Chapelle和Vapnik[3]基于支持向量的支撑(span)和特征空间的重构(rescaling)对SVM的模型选择问题进行了研究;Claeskens[4]给出了一种SVM模型选择的信息准则;Debruyne等[5]采用影响函数(influence function)对基于核函数的回归方法进行模型选择研究;黄景涛等[6-7]采用遗传算法和实验设计方法分别对SVM的参数选择进行了研究;还有不少学者结合应用背景采用各种启发式方法对SVM的参数进行选择研究。但这些研究往往只限于对算法分类正确率的优化,而没有特别考虑算法的运行速度问题。该文综合考虑SVM的分类正确率 r 和训练时间 t 两个指标对SVM的模型选择问题开展研究,采用多目标优化方法进行SVM的模型选择。

2 SVM算法原理及性能指标

2.1 SVM原理

SVM最初作为线性分类器提出,对于给定观测样本集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$,其中每个观测样本 $x_i=(x_{i1},x_{i2},\cdots,x_{id})$ 具有类别标签 y_i,对于两类问题 $y_i=\pm1$,SVM的目标是寻找一个函数(分类器),使得对于任一个观测样本,给出其对应的类别。

基金项目: 河南科技大学青年科学基金(No.QN07041)。
作者简介: 黄景涛(1977-),男,博士,副教授,主要研究领域为统计学习理论及其应用;池小梅(1979-),女,助理工程师,主要研究领域为模式识别及智能系统;马建伟(196-),男,博士,副教授,主要研究领域为系统控制与优化。
收稿日期: 2009-04-07 **修回日期:** 2009-06-15

若选择 0-1 损失函数,即:

$$V(y_i,D(\boldsymbol{x}))=\begin{cases}0, y_i=\mathrm{sgn}(D(\boldsymbol{x}))\\1, y_i\neq\mathrm{sgn}(D(\boldsymbol{x}))\end{cases}$$

SVM 训练过程就是使得损失函数最小化的过程,可转化为如下二次优化问题[1]:

$$\min_{\xi,c\in\boldsymbol{R}^l} C\sum_{i=1}^{l}\xi_i+\frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i(\boldsymbol{w}\cdot\boldsymbol{x}+\boldsymbol{b})\geqslant 1-\xi_i,\ i=1,2,\cdots,l \quad (1)$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,l$$

其对偶优化问题为:

$$\max \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j Q_{ij}$$
$$\text{s.t. } \sum_{i=1}^{l}y_i\alpha_i=0 \quad (2)$$
$$0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,l$$

其中,$Q_{ij}=y_iy_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$,$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$是满足 Mercer 条件的核函数;$C$ 是一个惩罚系数。通过求解问题(2)构建 SVM 分类器:

$$f(\boldsymbol{x})=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{x},\boldsymbol{x}_i)+b\right) \quad (3)$$

2.2 SVM 的性能指标

对于分类问题,算法的性能指标往往是指算法的分类正确率,即所构造的分类器在未知样本上的分类能力。大多研究都只关心分类正确率这一指标,而对分类器的训练速度关注相对较少。分类正确率无疑是衡量分类器性能的主要指标,但分类器的速度也是需要关注的,过分追求分类正确率在一些场合并不合适。对于一类分类器,其正确率达到一定程度之后再提升会很困难,小的正确率提升要牺牲比较大的训练效率。因此,在评价一个分类器性能时需要综合考虑分类正确率和训练速度。

对于 SVM 来说,训练正确率只需要将得到的分类器在测试样本集上进行检验即可。训练速度则可从算法复杂度和实际耗费时间两方面进行考察。对于 SVM 的模型选择来说,针对的往往是同一类别的分类器,在数学形式上属于同一类型的函数,其算法时间复杂度大致相当,该文从算法实际耗时方面考察算法的速度。从分类正确率和训练速度两方面考虑 SVM 的性能,需要对分类正确率和训练时间两个指标同时进行优化,模型选择就成为一个典型的多目标优化问题。

3 基于 PSO 的 SVM 多目标模型选择

3.1 SVM 算法的多目标性能分析

算法性能不同方面的权衡是一个典型的多目标优化问题,且各目标往往相互矛盾,不能同时达到最优值,这就需要在不同的目标之间进行折中。从分类正确率和训练时间两个方面综合考察 SVM 的性能,模型选择问题就成为一个多目标优化问题,即在 SVM 模型参数的可行集 $\boldsymbol{P}$ 内对分类正确率 r 和训练时间 t 进行同时优化,为便于求解,将正确率 r 最大化转换为最小化问题,即需要求解如下多目标优化问题:

$$\text{obj1: } \min\ -r;\quad \text{obj2: } \min\ t$$
$$\text{s.t.}\quad C,\gamma\in P \quad (4)$$

通过求解多目标问题(4)得到一个 Pareto 最优解集,然后根据具体情况选择相应的解作为最优解。对于 C-SVC 算法来说,如果将该算法用于在线学习或者其他对速度要求较高的场合,则从 Pareto 解集中选择耗时短的解,而分类精度则根据具体情况设定一个最低阈值。如果将该算法用于对分类精度要求很高、错分损失很大的情况,则着重考虑分类精度高的解。如图 1 所示,点 A 所代表的解精度最高,但耗时也最多,点 B 耗时最小,但精度也最低,点 C 则是二者的折中,实际应用时根据情况从 Pareto 解集中选择符合需要的模型,避免了改变应用场合时需要重新优化计算的问题。同时可以通过 Pareto 解集中解的分布情况来判断单个目标的变化趋势,在选择最优解时以某个目标较小的牺牲换取另一个目标较大的性能提升,使得总效益最大化。在 A、B 点附近,一个目标较小的损失就可以换取另一目标较大的提升,在 A 点附近,分类准确率的降低能够使得时间耗费量有较大幅度减小,因此即便是在对分类精度要求较高的场合,也不一定要选择点 A,可以考虑两个目标的综合情况而选择一个 A 点附近的点。同样,在 B 点附近,时间耗费量的增大能够使得分类准确率有较大提升,在对算法学习速度要求较高的场合可以选择 B 点附近的点而不是 B 点。

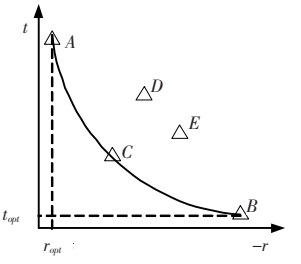

图 1　SVM 算法多目标 Pareto 解集示意图

3.2 SVM 多目标模型选择的求解

多目标优化问题的目标是一个向量空间,可以视为变量空间到目标空间的映射,其解是一个解集而不是一个唯一的最优解。多目标优化问题的求解方法有两类:将多目标优化问题转化为单目标优化问题进行求解和直接对多个目标同时进行优化。直接求解多目标优化问题的方法多为智能优化算法,可同时求出多目标规划问题的 Pareto 最优解集[8]。该文采用粒子群算法(Particle Swarm Optimization,PSO)对 SVM 模型选择问题进行求解[8]。粒子群算法是一种模仿自然界群体行为智能的算法,对各个粒子采用速度和加速度两个量确定其位置,每个粒子在每个位置对应一个适应度函数值,根据适应度函数值的大小来更新该粒子的速度,进而更新粒子的位置。这些粒子会根据自身的适应度调整位置,最终聚积在最优值附近。

PSO 算法能够以较快的速度逼近多目标优化问题(4)的 Pareto 最优解集。Pareto 解集的搜索过程是对 PSO 算法中各个粒子所代表的解的对比过程,从中选择支配解[8],当集合中所有解都为支配解时该集合就是 Pareto 解集。图 1 中的点 A、B、C 相互之间没有支配和被支配的关系,即这些点是一样“好”的点,点 D、E 则是被支配点。Pareto 解集是通过在可行解中进行支配关系的判断来获得的,算法流程如下:

步骤 1　初始化 PSO 参数;

步骤 2　产生一组可行解(初始化 PSO 群体);

步骤 3　按照支配关系进行比较,更新 Pareto 解集;

步骤 4　是否达到迭代次数?若是则转步骤 6;否则继续步骤 5;

步骤 5　更新粒子,得到下一代群体,转步骤 2;

步骤 6　结束,得到 Pareto 解集的一个近似。

4　仿真实验

采用上述方法,在几个数据集上进行了数值实验。测试平台为 PC 机,CPU 为 P4 2.0 G,768 M DDR 内存,Windows 2000 操作系统。SVM 的训练和测试采用 LIBSVM 算法包[9]。

图 2 是在 Banana 数据集[10]上进行优化的结果,图中横坐标是支持向量机算法的分类精度,纵坐标是相应参数下算法实际耗费的时间。分类精度和耗费时间是两个需要同时优化的目标,图中各点代表一组参数所对应的一个二目标值。Pareto 解集中包含了 8 个 Pareto 最优解,这些解对应的目标值具有一定的分散度,目标值中分类精度这个分量是离散的,这是因为数据集本身的原因造成的,该数据集包含了 400 个样本,分类精度采用的是 5-fold 交叉验证精度,所以每错分一个样本引起分类精度的变化是不连续的,当样本数足够多时,就趋向于连续分布。

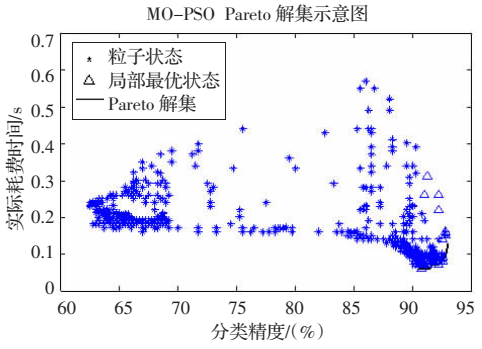

图 2　数据集 Banana 上的 MO-PSO 优化结果

图 3 是数据集 Diabetes[10]上的 Pareto 解集示意图。该数据集包含两类样本,每个样本有 8 个属性,一共 768 个样本。该数据集上采用 PSO 算法在参数空间内进行搜索得到的结果比较集中,不同的参数组合对应于一个目标点,粒子群在搜索过程中没有明显的轨迹,只得到了 4 个 Pareto 最优解。由于数据集较小,这些 Pareto 解之间两个目标之间的差别不大。

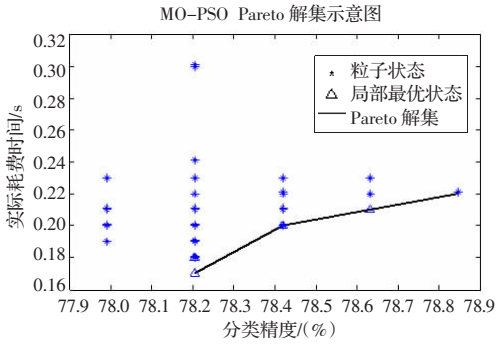

图 3　数据集 Diabetes 上的 MO-PSO 优化结果

图 4 是数据集 Image[10]上的寻优过程示意图,数据集 Image 也是一个二类分类问题,共有 1 300 个样本,每个样本有 18 个属性。图中给出了搜索空间中的可行解对应的两个目标值的情况,因为横坐标代表分类精度,需要最大化,而纵坐标代表耗费时间,需要最小化,所以 Pareto 边界相对于可行解对应的目标值位于右下角。在该数据集上进行参数多目标优化过程中,得到的 Pareto 最优解较多,表明在该数据集上不同参数的 SVM 分类效果和耗费时间有较大不同,这种情况更适合用多目标规划的方法进行模型选择。

图 5 是在数据集 a2a 上的目标值分布情况。由图可见,相

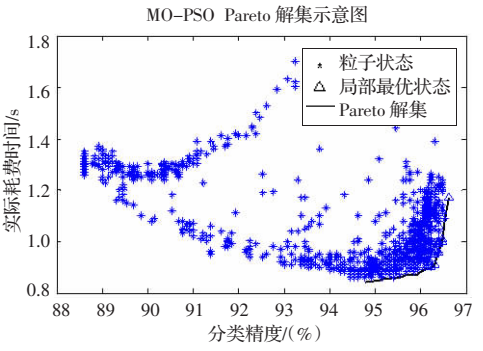

图 4　数据集 Image 上的 MO-PSO 优化结果

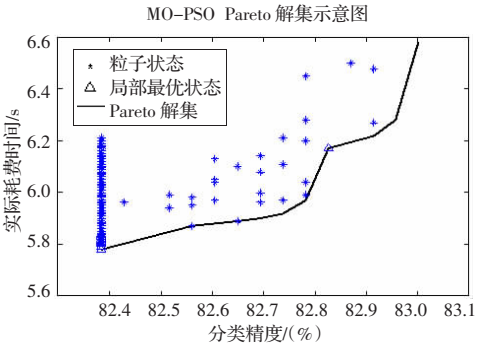

图 5　数据集 a2a 上的 MO-PSO 优化结果

当一部分目标值中的分目标分类精度集中在一个小的区间内,Pareto 解的分布较广泛。这些解对应的两个目标分量相差较大,在实际应用时根据具体要求进行选择。在该数据集上,所有 Pareto 解对应的参数 γ 都是同一个值,表明在该数据集上,参数 γ 对目标值的影响作用不如参数 C 的影响明显。

5　结论

从分类精度和训练时间两个方面对 SVM 分类器的模型选择问题进行了研究,结合多目标优化的思想,用 PSO 算法对多目标优化问题进行求解,在几个数据集上对 SVM 多目标模型选择进行了实验分析。实验结果中目标值的分布情况各不相同,这是由于数据集本身特性的不同引起的。Pareto 解集给出了 SVM 具体模型及其对应的性能指标,在具体选择 SVM 模型时,可以结合其他信息从 Pareto 解集中进行选择。多目标优化在很多决策问题中都存在,所提供的方法为 SVM 模型选择提供了一种新的思路。

参考文献:

[1] Hastie T J,Tibshirani R J,Friedman J.The elements of statistical learning:Data mining,inference,and prediction[M].Heidelberg:Springer-Verlag,2001.

[2] Vapnik V.The nature of statistical learning theory[M].New York,USA:Springer,1995:23-60.

[3] Chapelle O,Vapnik V.Model selection for support vector machines[C]//NIPS Conference,Denver,Colorado,USA,1999:230-236.

[4] Claeskens G,Croux C,Van Kerckhoven J.An information criterion for variable selection in support vector machines[J].Journal of Machine Learning Research,2008,9:541-558.

[5] Debruyne M,Hubert M,Suykens J A K.Model selection in kernel based regression using the influence function[J].Journal of Machine Learning Research,2008,9:2377-2400.

(下转 203 页)