利用粒子群算法缩减大规模数据集SVM训练样本

曾联明[1,2,3] 吴湘滨[1] 刘 鹏[3]
(中南大学地学与环境工程学院 长沙 410083)[1] (佛山科学技术学院信息中心 佛山 528000)[2]
(解放军理工大学网格技术研究中心 南京 210007)[3]

摘 要 对于大规模SVM训练样本数据,在分类前采用粒子群算法进行样本缩减,每一个粒子的维对应一个样本状态,通过更新粒子的速度和位置信息,调整训练样本的状态,引导粒子向分类最优的样本状态组合方向移动,去除样本中对分类不起作用的非支持向量和冗余的支持向量所对应的样本,生成新的缩减样本,进行分类训练,从而达到提高训练效率的目的。基于大规模遥感图像数据集的分类实验表明,此方法在确保不降低分类精度的前提下减少了分类时间。

关键词 粒子群,支持向量机,训练样本,海量数据
中图法分类号 TP315 **文献标识码** A

Sample Reduction Strategy for SVM Large-scale Training Data Set Using PSO

ZENG Lian-ming[1,2,3] WU Xiang-bin[1] LIU Peng[3]
(School of Geosciences and Environmental Engineering, Center South University, Changsha 410083, China)[1]
(Information and Educational Technology Center, Foshan University, Foshan 528000, China)[2]
(Grid Technology Research Center, The PLA University of Science and Technology, Jiangsu 210007, China)[3]

Abstract A PSO algorithm reduction strategy was proposed to a SVM large-scale training samples by updating the velocity and location of the particles, each particle was corresponding to the status of the training samples, the ideal status included the smallest number of sv, the new training sample has reduced some nsv which are not effect the SVM classification, so as to reduce the size of the training data sets. A practice of the remote session image classification has proved that the strategy not only has reduced samples, but also enhanced the efficiency of the large-scale data sets training.

Keywords PSO, SVM, Training sample, Large-scale data

1 引言

SVM(Support Vector Machines,支持向量机)由于较好地解决了小样本、非线性、高维数、局部极小点等实际问题,已经逐渐成为解决模式分类问题的首选工具[1]。但对于大规模数据的分类训练问题,SVM却遇到了计算量庞大的现实难题。因为SVM是一个凸二次规划问题,在求解的过程中,虽然保证了解的存在和唯一,但训练样本的个数直接决定问题求解的维数,从而左右运算的规模,个数越多,规模越大[2]。基于SVM的解具有稀疏性的特点,SVM样本中对分类有贡献的点称为支持向量,没有贡献的点称为非支持向量,文献[3]证明了非支持向量机不仅与支持向量的决策无关,也不会影响支持向量机的训练过程和训练结果,支持向量只是很少的一部分,非支持向量在问题求解的过程中占用了很大一部分计算资源,训练样本数据集规模越大,此问题变得越突出。研究如何在训练前去除训练样本中非支持向量的样本,缩减无效样本数量,节约运算和传输资源,提高训练效率,就变得非常有意义。

PSO(Particle Swarm Optimization,粒子群算法)是由Kennedy和Eberhart受鸟觅食行为启发于1995年提出的一种基于群体智能理论的随机寻优算法[4]。PSO具有原理简单、通用性强、支持群体搜索和协同搜索、鲁棒性好等优点,在神经元网络训练、化工和电力系统、图像处理、生命信息、机械设计等领域已得到有效应用[5]。

本文采用粒子群智能算法,通过粒子的飞行速度和位置的更新随机组合训练样本的状态,以是否满足分类精度的要求为决策条件,求解训练样本的最小化组合,以达到样本缩减的目的。

2 SVM分类与PSO算法

2.1 SVM分类流程

支持向量机的分类过程就是求解最优化的过程。对于线性问题和非线性问题,其在分类中的应用,只需要将输入向量非线性映射到一个更高维的特征空间,然后构造最优分类超平面[6]。SVM的分类流程,通常包括样本采集、样本格式转换、样本训练、样本预测和分类几个部分。图1所示为增加样

到稿日期:2008-10-27 返修日期:2009-01-07 本文受国家自然科学基金(40473029)资助。
曾联明(1973-),男,博士研究生,主要研究方向为网格计算技术、遥感图像分类与信息集成,E-mail:zenglm@163.com;吴湘滨(1962-),男,教授,主要研究方向为地质灾害评价、预测与防治、地理信息;刘 鹏(1970-),男,教授,主要研究方向为高性能计算、网格技术。

本缩减处理环节后的 SVM 分类流程。

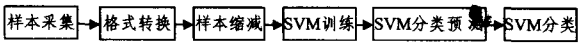

图 1　增加样本缩减处理后的 SVM 分类流程

对样本的缩减处理在样本训练之前进行，由于决策的需要，缩减处理的过程也涉及到不断的精度计算。

2.2　基本粒子群算法

粒子群算法将问题的搜索空间类比于鸟类的飞行空间，将每只鸟抽象为一个无质量、无体积的微粒，用于表示待求解问题的一个候选解，解的最优化所得就等同于要寻找的食物。设群中的粒子总数为 N，待搜索问题的解空间为 D 维空间。向量 $X_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(D)})$ 表示第 i 个粒子位置，$P_i=(p_i^{(1)},p_i^{(2)},\cdots,p_i^{(D)})$ 为该粒子已搜索到的个体最优位置，$P_g=(p_g^{(1)},p_g^{(2)},\cdots,p_g^{(D)})$ 为粒子群已搜索到的群最优位置，第 i 个粒子的飞行速度为 $V_i=(v_i^{(1)},v_i^{(2)},\cdots,v_i^{(D)})$。每个粒子的速度和位置按式(1)计算[7]：

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

其中，ω 称惯性因子，c_1 和 c_2 为正的加速常数，φ_1 和 φ_2 为 0 到 1 之间均匀分布的随机数。ω 较大，适合对解空间进行大范围探查；ω 较小，适合进行小范围搜索。另外，为了对微粒的移动进行适当的限制，可以设置第 d 维的位置变化范围 $[-XMAX_d,XMAX_d]$ 和速度区间 $[-VMAX_d,VMAX_d]$。

PSO 算法具有记忆微粒最佳位置的能力和微粒间信息共享的机制，即通过种群间个体的合作与竞争来实现对优化问题的求解。

3　SVM 训练样本 PSO 缩减流程

3.1　规则构造和适应度的定义

对于任意的经过预处理(格式转换、归一化)后的训练样本，通过规则构造模仿鸟类觅食的行为，粒子在最优上界和下界所构成的维度空间里搜索最优值，对所有维度微粒状态的最优组合的求解过程形成一条规则。

将参与训练的整个样本数据文件作为种群的一个粒子，一个种群可以包含多个粒子，每个粒子单独进行飞行，所有的粒子遵循相同行为规则，粒子与粒子之间相互共享信息。样本文件中的每个样本作为粒子的一维，称为一个微粒。微粒的状态由数组 $Status[Pdimension]$ 来存放，$Pdimension$ 是维的宽度。每维对应一个变量 $Status[i]$，其值为 0 或 1，用于表征该样本是否为支持向量。$Status[i]=1$ 表示该维为 SV(支持向量)，参与样本训练 $Status[i]=0$ 代表该维为 nSV(非支持向量)，不参与样本训练。每个粒子就是由状态为 $Status[i]$ 的 D 维样本组成，D 为样本总数，$0<i\leqslant D$。粒子飞行的过程中，每个粒子中微粒的状态是随机生成的序列，由 0 和 1 构成，如 $Status[D]=\{0,1,1,0,1,\cdots,0\}$ 就是一个粒子某一时段的一个状态。

根据微粒飞行的速度、个体最优和群最优以及权重来调整微粒飞行的位置，按式(1)进行计算，然后将位置的值跟一个阈值进行比较。如果位置的值大于等于该阈值，则设置该微粒的状态值为 1，否则为 0。判断随机飞行的粒子微粒组合是否达到最佳，其依据为新组合下的样本分类精度。该分类精度由 K-折交验证来得到。每次 K-折交验证必须使用相同的超参数，以确保分类精度的可比性，只有精度不降低条件下的微粒组合才是有效缩减样本组合。

适应度定义为粒子的一组微粒状态组合，适应度的值用分类精度来进行衡量。每个粒子每飞行一次，都生成新样本，都要计算一次适应度的值。如果新样本的适应度值比当前最优适应度值好，则将当前适应度设置为当前最优，否则继续飞行，直到一次迭代的所有粒子飞行完为止。粒子每飞行一次都需要更新一次群最优适应度，并将更新结果传递到下一次的迭代。当迭代次数超过最大迭代次数时，循环终止。

3.2　缩减流程

基于 PSO 算法的 SVM 训练样本缩减流程如图 2 所示。

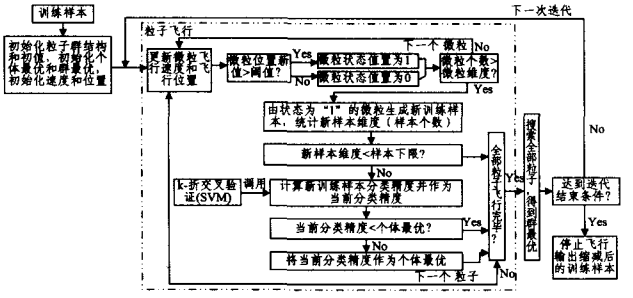

图 2　基于 PSO 算法的 SVM 训练样本缩减流程

首先，设置种群大小，此处只研究单种群的最优问题，其值设为 1。

其次，初始化粒子参数和状态值，这些参数包括粒子数、最大迭代数、微粒状态初值、个体微粒最优和群微粒最优、最优适应度和最优群适应度。在粒子飞行之前，其每个微粒状态初值均为 1，即假设全部微粒都是 SV，参与样本训练，群最优和个体最优相等，则个体适应度的值等于群最优适应度的值。

接下来粒子进行飞行，直至达到最大迭代次数为止。粒子飞行过程中，随机生成由 0 和 1 构成的微粒状态序列。缩减后新的训练样本根据微粒的状态来生成，状态为 1 的微粒(样本)被写入新的样本文件，状态为 0 的微粒被缩减掉，不被写入。

最终，所有迭代完成后得到的群最优状态组合就是经过 PSO 样本缩减后的样本最优组合，这些缩减后的样本将被作为正式参加样本训练的样本。

在整个缩减流程的粒子飞行过程里，决定微粒的状态为 1 还是 0，阈值如何获得显得非常重要。本文中阈值由所有微粒的非零位置值通过归一化后的均值来得到。这样处理可以避免粒子的随机状态受大数据的影响而陷入局部收敛。

当全部迭代完成后，需要根据缩减结果即根据群最优微粒组合的值与原样本数据进行比对，状态值为 0 的样本统统被舍去，不参与 SVM 训练。

4　实验分析

基于以上的设计和分析，将 PSO 群智能缩减策略应用到遥感图像的 SVM 分类中。图 3 为实验所用的原始唐山遥感 TM 图像，其分辨率为 1∶50000，图像大小为 500×318，样本总数为 3756 个。实验环境：windows 操作系统，编程语言 Visual Studioo 2005 C#，SVM 开源程序 Libsvm 2.6，遥感软

件 ENVI 4.1。试验用计算机硬件配置：Intel P4 CPU 1.8GHz，512M 内存。

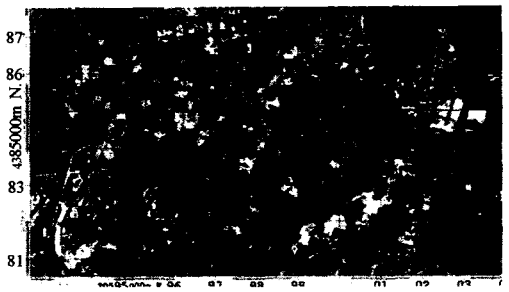

图 3　唐山地区 1∶50000TM 局部图(黑色部分为水面)

本文试验使用多类问题的分类求解。设训练样本集为 $(x_1,y_1),\cdots,(x_l,y_l),x_i\in R^n$，$x_i$ 为训练样本输入，$i=1,\cdots n$，y_i 为期望输出，表示类别标号。对于本文所讨论的问题，采用 C-SVM 分类方法，核函数为径向基核函数，$c=128$，$\gamma=0.125$。共有 4 类地物，分别表示房屋、水面、绿地和公共用地，即 $y_i\in\{4,3,2,1\}$。训练样本通过遥感软件 ENVI 的 ROIs 工具来获取。

PSO 中基本参数设置为 $c_1=2,c_2=2,\omega=1$，粒子群数量为 15，最大迭代次数为 20。PSO 算法运行的结束条件是达到最大迭代次数即刻停止。表 1 所列为训练样本经本文所述策略缩减前后分类相关参数的比较。

表 1　SVM 训练样本缩减前后分类指数比较

	训练样本总数	缩减率	最优分类精度	总耗时(秒)	支持向量(sv)个数				
缩减前	3756	0	83%	41	房屋 33	水面 608	绿地 601	公共 78	合计 1320
缩减后	1516	59%	84%	20	17	237	237	46	537

图 4 所示为样本缩减前的 SVM 分类结果-水面(蓝色部分)，图 5 所示为缩减样本后的试验结果。结合表 1 可知，基于 PSO 的缩减策略，不但丢弃了非支持向量数部分，也去除了训练样本中大量冗余的支持向量，使得训练样本得到了很大的缩减，训练和分类总消耗时间缩短了大约一半，其分类精度提高了 1%。

图 4　没缩减样本的分类结果-水面(黑色部分)

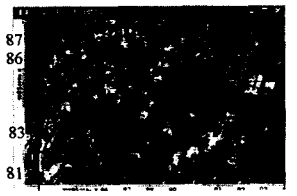

图 5　缩减样本后的分类结果-水面(黑色部分)

结束语　基于 SVM 的图像分类已经得到了广泛的应用，但也面临着挑战，特别是对大规模数据集样本的训练依然非常困难，本文采用粒子群算法对训练样本进行优化缩减，尽可能地将训练样本中非支持向量和冗余的支持向量去除，以提高训练效率。通过对多类遥感图像的分类试验证明，此方法在保证分类精度的前提下，达到了缩减的目的，这对于大规模数据集支持向量样本的训练效率的提高有非常重要的现实意义。如何进一步提高分类精度，是下一步需要研究的内容。

参 考 文 献

[1] Zhang Xue gong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32-42

[2] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines [M]. Cambridge, UK: Cambridge University Press, 2000

[3] 罗瑜，易文德，王丹琛，等. 大规模数据集下支持向量机训练样本的缩减策略[J]. 计算机科学，2007，34(10)：211-213

[4] Kennedy J, Eberhart R C, Shi Y. Swarm Intelligence[M]. San Francisco: Morgan Kaufman Publishers, 2001

[5] 王凌，刘波. 微粒群优化与调度算法[M]. 北京：清华大学出版社，2008

[6] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. House of Electronics Industry, 2005

[7] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[C]// Proceedings of the Word Multiconference on Systemics, Cybernetics and Informatics. Piscataway, NJ, 1997

(上接第 204 页)

[4] 欧阳为民，蔡庆生. 在数据库中发现具有时态约束的关联规则[J]，软件学报，1999，10(5)：527-532

[5] 欧阳为民，蔡庆生. 数据库中的时态数据发掘研究[J]. 计算机科学，1998，25(4)：60-63

[6] 徐敏，金远平. 一种新的周期性关联规则模型[J]. 计算机工程与科学，2000，22(4)：78-81

[7] Garofalakis M. Mining sequential patterns with regular expression constraints[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 120-136

[8] Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding Association Rules by using Confidence and Support[C]// Proceedings of the 4th Information Hiding Workshop. 2001: 369-383

[9] Pinkas B. Cryptographic techniques for privacy - preserving data mining[J]. SIGKDD Explorations, 2002, 4(2)

[10] Chang Liwu, Moskowitz I S. Parsimonious downgrading and decisions trees applied to the inference problem[C]// Proceedings of the 1998 New Security Paradigms Workshop. 1998: 82-89

[11] Evfimievski A, Srikant R, Agrawal R, et al. Privacy preserving mining of association rules[C]// Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 2002: 217-228

[12] Kantarcioglu M, Clifton C. Privacy- preserving Distributed Mining of Association Rules on Horizontally Partitioned Data[C]// ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. 2002

[13] Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data[C]// the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 639-644

[14] Oliveria S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining[C]// Workshop on Privacy, Security and Data Mining at The 2002 IEEE International Conference on Data Mining (ICDM'02). Maebashi City, Japan, December 2002