文章编号:1006-9348(2008)02-0096-04

一种基于感知器的样本空间划分方法

丛 聃,吕宝粮

(上海交通大学计算机科学与工程系,上海 200240)

摘要:二类分类问题是机器学习中的最基本的一类重要问题。目前广泛使用的,也是最为有效的学习算法是支持向量机(SVM)。然而对于某些非线性分类问题,SVM 还不能给出令人满意的解,因此希望能找到一种方法对 SVM 解决非线性分类问题的能力加以改进。对二类分类问题,提出一种基于感知器的样本空间划分方法。该方法首先用感知器提取样本的分布信息,将整体问题划分为局部空间中的分类问题,而后使用 SVM 求出各个局部问题的最优分界面,并用最小最大模块化网络对局部分界面进行综合,得到问题的全局解。仿真实验表明,新方法能够有效地分析样本空间,提取样本分布信息,在测试数据上得到了比原有方法更好的准确率。新方法实现了预期的目标,提高了分类器处理非线性分类问题的能力。

关键词:感知器;支持向量机;模式识别;样本空间分析;最小最大模块化网络

中图分类号:TP391 **文献标识码**:A

Partition of Sample Space with Perceptrons

CONG Chong, LU Bao-Liang

(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai, 200240, China)

ABSTRACT: Binary classification is a fundamental problem in machine learning area. Currently, the widely used and best performing learning method is support vector machine (SVM). However, SVM cannot give satisfactory answer to some non-linear problems. A new approach to improve SVM's ability on non-linear problems is expected. Focusing on binary classification problems, a novel sample space analyzing method based on perceptron is proposed. The method starts from extracting sample distribution information, and divides the overall problem into a series of local problems. Then finding optimized local separator with SVM, and finally combining the local separators with the minimization and the maximization principles to get the overall classifier. The simulation results indicate that the proposed method can effectively analyze sample space and extract distribution information, and achieve better prediction accuracy than existing methods. The new method can meet the requirement, and improve classifier's performance on non-linear problems.

KEYWORDS: Perceptron; Support vector machine (SVM); Pattern recognition; Sample space partition; Min-max modular network

1 引言

自动分类问题是机器学习的重要研究课题之一。最典型的分类问题是二类分类问题。该问题给出一组数据 $X=\{x_i:y_i\}_{i=1}^{N}$,其中 x_i 属于 d 维实数空间 R^d, y_i 属于集合 $\{0,1\}$,其中 y_i 为 0 的样本称为负样本,y_i 为 1 的称为正样本,N 为训练集样本数量。要求学习算法总结这组数据的规律,并构造出一个分类器,利用学习到的算法,对训练集中未出现过的新样本进行分类。学习过程中出现的样本统称为训练集,训练中未出现过的样本可用于测试分类器的效果,统称为测试集。

大量文献表明,支持向量机(SVM)为机器学习中的二类分类问题,提供了绝佳的分界面搜寻算法。然而在某些问题上应用 SVM 时,得到的结果仍然不甚理想,有相当的改进空间。

经过对 SVM 基本理论的分析,发现 SVM 解决分类问题时隐含一个假设,即样本在其空间中的分布应该较为简单,甚至线性可分。但有时这一假设并不成立。

最小最大模块化网络(Min-Max Modular Network, M^3)[1],通过对训练样本的划分和精心设计的组合规则,实现了

基金项目:国家自然科学基金(60375022, 60473040)和上海交通大学微软智能计算与智能系统实验室的资助

收稿日期:2007-01-26 修回日期:2007-01-31

将一个较大的分类问题划分为多个相对独立的小分类问题的目标。

通过对 M^3 框架与 SVM 各自优势的分析，本文提出了一种解决分类问题的新方法。首先对训练样本的预学习以确定合适的样本空间划分规则，按照训练样本的分布特点，将复杂的分类问题分解为多个较为简单的局部分类问题；然后用 SVM 求出这些局部问题的解；最后用 M^3 框架综合各个局部分类器。新方法在划分过程中通过保留和提取样本的分布信息，实现了对问题空间更有效的分析，提高了分类精度。

2 直接应用支持向量机的局限性

SVM 的核心思想是最大分类边界理论，该理论要求所求解的二类分类问题线性可分。许多实际问题显然不符合这一要求，为解决这一问题，SVM 在这一核心思想上做了两点补充：

一是利用软边界（Soft Margin），对越过边界的样本点计算一定的惩罚，而不是完全禁止。即使训练样本中存在少量的噪声，使一个线性可分问题变得不可分，使用软边界的 SVM 也能给出合理的分类边界。

如果正负样本点的分布本身是线性不可分的。那么即使利用软边界，也无法得到合适的分类边界。SVM 通过使用核函数，将原空间中的样本点映射到更高维空间上。由于维数的增加，一些原本不可分的问题变得线性可分。

向高维映射的方法有两个潜在的问题，一是训练样本点映射到高维空间后，是否能够反映它们在原样本空间中的分布；二是在高维空间中最大化分类边界而得到的最佳分类面，是否在原空间中也是最佳的。

由于高维空间很难给人以直观的印象，上述两个问题难以找到简洁直观的答案。如果可以在原样本空间中，通过分析样本的分布，把复杂问题转化为一系列较为简单的问题，并用这些问题的解，即一系列分界面，去描述原问题分界面，可以对复杂的分类问题得到更好的效果。

3 最小最大模块化框架

最小最大模块化（M^3）框架是将问题分而治之的有效方法。该方法可分解训练样本集，从而增加机器学习程序的并行性。除此之外，只要能考虑样本在空间中的分布，通过对训练样本进行预处理，就可以用 M^3 方法分析样本空间，从而提高分类器的性能。

若存在训练集 X，包含有正负两类训练样本。M^3 划分过程接收一个整数参数 L，作为子问题大小的指导标准，将正负样本集分别划分为 m，n 个子集。例如图 1 所示，将正样本分为 2 组（白色方块），负样本分为 3 组（黑色方块）。每一组正样本与每一组负样本分别组合，共可以得到 6 个二类子问题。分别学习这些子问题，再将得到的 6 个分类器按照图 2 所示的方法结合起来，就得到整个问题的解。

由于希望能够分析训练样本的分布规律，并在正负样本交界处找到最佳分界面。因此按照 M^3 框架划分训练样本时应该尽可能满足如下两点要求：

首先，要保留样本在其空间中的分布信息，以便为 M^3 框架中每个子分类器提供尽可能多的信息；其次，应使子问题的样本在空间中形成一个尽可能简单的二类分类问题，以利于 SVM 在子问题的学习中发挥最佳效果。

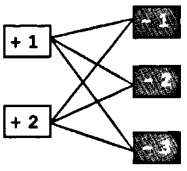


图 1 训练样本划分

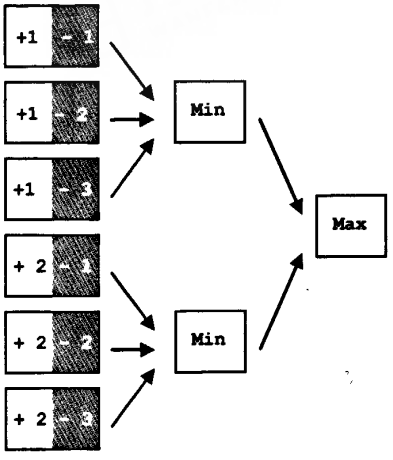


图 2 用 M3 框架组合局部分类器

为达到这两点要求，在划分时需要综合考虑正负样本，统一进行划分，待分组完成后再按图 1 的方式将正负样本完全分离。对于数据的分组，希望划分后同一组中的数据具有如下特性：

1）各个样本在空间上相互接近

2）信息熵较小（即二类样本所占比例差别大）

另外子问题的平均粒度不宜过小，以免学习出的整个模型过于复杂，引起过度拟合。

按上述要求划分训练样本而得到的子问题，能够在相当的程度上完整地描述样本在其空间上的分布情况。这时再使用 SVM 寻找每个子问题的最佳分界面，可以使整个问题的解更加优化。

4 样本空间划分算法 M^3 - perceptron

要使同组内样本在空间上聚拢，可以利用比较规则的分界面划分训练数据。要使同一组中的样本比例差别大，应该尽量将不同类型的数据置于分界面两侧。综合两项考虑，选择了感知器（perceptron）[2] 作为数据划分标准。因为用感知器能得到非常规则的线性分界面，而且感知器倾向于将不同类样本划分到不同组中，使同一组内样本纯度比较高。

算法采用分而治之的策略：首先将所有训练数据看作一组，用感知器学习这组数据，并用学习出的结果将这组数据

划分为两组;之后在这两组数据上递归地重复这一过程,直到分出的小组中正样本或负样本的数量少于预先指定的上限。算法流程如下:

输入:数据集 $X=\{x_i:y_i\}_{i=1}^{N}$,其中 $x_i \in R^d$,$y_i \in \{0,1\}$ 划分模块大小上限 L

初始化:数据分块集合 $B=\{X\}$

算法:

```
function size(c)
    return min(c 中的正样本数,c 中的负样本数)
end

while true
    在 B 中找到 c,使得 size(c) 最大
    if size(c) < = L then break
    用 c 中的样本训练感知器 p
    用 p 划分 c 中的样本,得到 c1, c2
    从 B 中移除 c
    将 c1, c2 加入 B
end
```

感知器的训练是一个迭代过程,为避免长时间等待,设置了迭代次数上限,因此感知器学习全部训练数据的时间为 $\Theta(N)$,其中 N 为训练样本的个数。本算法与快速排序算法(quick sort)非常相似,可用类似的方法估计时间复杂度。与快速排序不同的是,当数据被划分为不超过 L 大小的组时,递归立即停止。综上,整个算法的时间复杂度估计为 $\Theta(N \cdot \log(\frac{N}{L}))$。

5 仿真实验

5.1 实验设置

为评估所提方法的性能,通过一系列实验将新方法与已有方法进行比较。第一个数据集是 Banana 问题[3],第二个数据集是 UCI 的字母识别问题[3]。Banana 数据集有训练、测试数据各 100 组,我们将其中前 5 组训练数据合并,作为实验的训练集,将第一组测试数据作为实验的测试集。字母识别问题首先按照文献[3]中的方法转化为一个二类问题,即"E","H","I","M","P","Q","R","X","Y","Z"标记为正样本,而其它类别全部标记为负样本。两个数据集的一些属性如表 1 所示:

表 1 数据集属性

数据集	样本空间维数	类别数	训练样本数	测试样本数
Banana	2	2	2000	4900
字母识别	16	2	15000	5000

本文采用 svm - light[4] 作为 SVM 实现。SVM 的参数 c(软边界算法对误分训练样本点的惩罚系数)由该程序自动选择,使用径向基函数(RBF)核时,参数 g 选择了 0.1, 0.2, 0.3, 04, 0.5, 0.6, 0.7, 0.8, 0.9 与 1.0 共 10 个值分别实验,并将最好的结果列出。为比较划分模块上限(算法参数 L)对泛化精度的影响,跟据训练样本的数目选择了一系列不同的 L 值(见下一小节)。

5.2 实验结果与讨论

下列表格中列出了各种方法在两个数据集上的测试精度,度量标准为正负样本 F - 1 值的微观平均值[5]。表格第一列是算法参数 L 的取值,该参数调整划分模块的大小。由于直接使用 SVM 时不需要划分数据,L 参数为原数据集大小。第二列是直接使用 SVM 学习得到的泛化精度。后三列都使用了 M^3 框架,但采取了不同的划分策略。其中 M^3 - random 和 M^3 - superplane[6] 分别采用了随机划分与超平面划分策略,M^3 - perceptron 采用了本文提出的策略。

表 2 Banana 数据,线性核

L	SVM	M^3 - random	M^3 - superplane	M^3 - perceptron
2000	0.4801	-	-	-
600	-	0.4823	0.7062	0.6649
500	-	0.2695	0.7743	0.8342
400	-	0.4934	0.7530	0.7896
300	-	0.4757	0.7263	0.8506
200	-	0.5692	0.8040	0.8462
100	-	0.5288	0.8184	0.8655

表 3 Banana 数据,径向基函数(RBF)核

L	SVM	M^3 - random	M^3 - superplane	M^3 - perceptron
2000	0.8981	-	-	-
600	-	0.8969	0.8992	0.8987
500	-	0.8949	0.8960	0.8980
400	-	0.8970	0.8980	0.8988
300	-	0.8957	0.8931	0.9005
200	-	0.8919	0.8926	0.8946
100	-	0.8908	0.8892	0.8957

表 4 字母识别数据,线性核

L	SVM	M^3 - random	M^3 - superplane	M^3 - perceptron
15000	0.4608	-	-	-
6000	-	0.5961	0.6236	0.6233
5000	-	0.4608	0.5617	0.6665
4000	-	0.6392	0.6299	0.6665
3000	-	0.5907	0.6255	0.6574
2000	-	0.6415	0.6277	0.6851
1000	-	0.6437	0.6491	0.7336

表5 字母识别数据,径向基函数(RBF)核

L	SVM	M^3 - random	M^3 - superplane	M^3 - perceptron
15000	0.8343	-	-	-
6000	-	0.8002	0.8321	0.8323
5000	-	0.8129	0.8293	0.8377
4000	-	0.8158	0.8110	0.8377
3000	-	0.7675	0.8167	0.8336
2000	-	0.7943	0.7965	0.8422
1000	-	0.7664	0.7837	0.8415

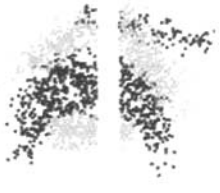
Banana 数据是典型的线性不可分问题,由表 2 可看出,分析和提取数据的分布信息,对泛化能力的提高起到了一定作用。图 3 与图 4 展示了 M^3 - perceptron(线性核)在 L = 500 时学习到的分界面。

图3 M3 - perceptron 学习到的分类边界

图4 分类边界与训练数据的对比

图 5 显示了 Banana 数据的分布情况。图 6 和图 7 显示了 L 取值为 500 时 M^3 - perceptron 划分训练数据的过程。根据参数 L,训练数据被划分为三部分。划分后将每部分正负样本单独提出,可得到正样本三组,负样本三组。其中位于同一部分的正负样本组成的分类子问题是最关键的局部问题。由图 7 发现,左边的两部分样本是接近线性可分的,右边的一部分,由于正负样本数均已小于 L,没有继续划分。总体上看,本文提出的方法可以准确分析样本在其空间的分部。如果 L 的取值更小,所有子问题都能达到接近可分的程度。

训练数据



第一次划分

图7 M^3 - perceptron 最后划分结果

另外从表格中还发现,随着子问题划分的细化,即随着 L 取值的减小,M^3 - perceptron 的泛化能力有一定的上升趋势。这与的期望相符,但是过细的划分可能导致过度拟合。因此如何为 L 选择合适的取值,以及这种上升趋势与 L 取值之间的具体关系,将是下一步研究的方向。

6 结论与展望

本文利用感知器算法对样本空间进行分析,结合支持向量机和最小最大模块化方法,提出了一种解决分类问题的新方法。新方法的特点是它应用了由粗到精,分而治之的思想,首先用简单快速的感知器算法,从全局上对样本的分布情况作初步分析,在此基础上利用 SVM 求出各个局部最佳分界面,并通过 M^3 框架将局部最佳分界面综合,从而得到全局的非线性分界面。实验中观察到,新方法抓住了样本在空间中的分布信息,获得了比原有方法更好的泛化预测能力。

算法参数 L 的选取在一定程度上影响模型的泛化能力,后续工作中希望找到一种自动选取 L 的方法,以使算法在使用上更加方便、性能上更加可靠。

参考文献:

[1] B L Lu and M Ito. Task decomposition and module combination based on class relations: a modular neural network for pattern classification[J]. IEEE Trans. Neural Networks, vol. 10, pp. 1244 - 1256, 1999.

[2] T M Mitchell. Machine Learning[M]. section 4.4. McGraw - Hill, 1997.

[3] G. Rätsch, T Onoda and K R Müller. Soft margins for AdaBoost [J]. Machine Learning, 2001, 42: 287 - 320.

[4] T Joachims. Making large - Scale SVM Learning Practical[M]. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT - Press, 1999.

(下转第 103 页)

表4 三种识别方法的结果图

训练样本数	测试样本数	RBF 神经网络识别率	支持向量机识别率	线性规划识别率
30	30	83.33%	90%	86.67%
60	30	90%	无法识别	93.33%

与其它识别方法相比较,线性规划有着较高的识别率(见表4)。可以从上文看出,线性规划类似于特征样本在转化后的空间内的区域聚类,同类特征样本属于一个聚类,而另一类特征样本属于另一聚类,这比较符合实际情况,即同类特征样本之间的距离短而且比较集中,而不同类的特征样本之间的距离长而且比较分散。另外,拟支持向量的提取较简单,没有要凭经验确定的未知数。而支持向量机识别的精度由惩罚因子和核参数两个未知数决定,要正好选中最好的是不可能的,因此这两个未知数影响了识别的精度,需要经验和大量的实验获得。

因此,由实验结果可以看出,本文提出的算法可以用来进行飞机目标识别,而且训练数目越大,识别的结果越好,同时识别的时间较短。但与支持向量机进行识别比较,可以知道训练和识别的时间明显少于支持向量机的识别,比如当有60个训练样本进行支持向量识别运算时,维数太大,计算机的内存远远不够,而且速度很慢。所以需要进行大量的样本训练时,而且要求速度要快时,支持向量机不能满足要求。采用本文的算法可以满足要求,而且识别的效果不错。

6 结论

本文根据空间变化,将原样本空间转化为另一空间,以便运用线性规划进行样本的拟支持向量的提取,再运用拟支持向量对应的判决向量进行识别,实验表明本文提出的算法可以进行飞机目标的识别,而且识别精度高。与其它识别方法相比较,线性规划识别率最高,而且所需要的内存少。特别与支持向量机相比较,拟支持向量的提取要比支持向量的提取简单,它没有未知数需要我们假设,所以识别的精度要高。同时支持向量机是运用的二次规划,这较线性规划复杂,所需的内存要大,如果样本数量大,计算机就无法运行。线性规划正好解决了这个难题。

参考文献:

[1] 陈东,王炎,崔有志. 基于旋转不变性小波矩的神经网络飞机识别[J]. 计算机工程与设计, 1998, 19(3):3-6.

[2] 王强,曹仲冬,马振晖. 一种新型 RBF 神经网络及其在舰船雷达目标识别中的应用[J]. 现代电子技术,2003, 149(6):95-98.

[3] 王长春,刘隆和. 多传感器目标的模糊识别[J]. 雷达科学与技术,2003, 1(2): 69-73.

[4] Lee Yuh-Jye, O L Mangasarian. SSVM:a smooth support vector machine for Classification[J]. Computational Optimization and Application,2001,20(1):5-22.

[5] András, Péter. The Equivalence of support vector Machine and Regularization Neural Networks[J]. Neural Processing Letters, 2002,15(2):97-104.

[6] Ming-Hsuan Yang, Narendra Ahuja. A Geometric Approach to Train Support Vector Machines. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. IEEE, 2000. 430-437.

[作者简介]

梅 蓉(1978-),女(汉族),江苏海安人,硕士研究生,助教,主要研究方向:图象处理与识别技术。

(上接第99页)

[5] F Sebastiani. Machine Learning in Automated Text Categorization [J]. ACM Computing Surveys, 2002, 34: 1-47.

[6] K A Wang, H Zhao and B L Lu. Task decomposition using geometric relation for min-max modular SVMs[C]. Lecture Notes in Computer Science,, 2005, 3496: 887-892.

[作者简介]

丛 翀(1981-),男(汉族),山东人。上海交通大学计算机系研究生,主要研究领域为大规模并行机器学习,以及自动文本分类问题。

吕宝粮(1960-),男(汉族),山东人。上海交通大学计算机系教授,博士生导师,IEEE 高级会员。主要研究领域有仿脑计算机理论与模型、计算系统生物学、脑-计算机接口以及自然语言处理。