

支持向量机分类算法在 MATLAB 环境下的实现

董 婷

(榆林学院 计算机与网络工程系, 陕西 榆林 719000)

摘 要:支持向量机算法 SVM(Support Vector Machine)作为新一代机器学习算法近年来被成功的应用到很多模式识别问题中,其在数学上表示为求解一个二次规划问题。主要论述了支持向量机分类算法在 MATLAB 环境下的具体实现方法,为支持向量机算法的学习者和非计算机专业的广大研究人员提供一种简单、方便、高效、可行实现方法。

关键词:SVM;二次规划;MATLAB

中图分类号:TP181 **文献标识码:**A **文章编号:**1008-3871(2008)04-0094-03

V. Vapnik 等人从二十世纪六、七十年代致力于小样本的机器学习研究,到二十世纪九十年代中期,统计学习理论受到越来越广泛的重视^[1],研究如何从一些观察数据(样本)出发,模拟目前为止尚不能通过原理或实验发现的规律;利用这些规律分析客观对象,对未来数据或无法观测的数据进行预测,这就是机器学习的统计方法^[2]。支持向量机 SVM(Support Vector Machine)是在统计学习理论基础上发展起来的一种新的机器学习方法,是结构风险最小化原理的实现^[3]。算法实现需具有深厚的数学功底和计算机编程技术,对非计算机专业的广大研究人员来说,一种简单高效的实现环境和方法是迫切的需要。支持向量机算法在 MATLAB 环境下易于实现和灵活应用的特点,很好的提供这一技术平台。

1 支持向量机及 MATLAB

1.1 最优超平面 SVM 方法是从线性可分的情况下的最优分类面(Optimal Hyperplane)提出的。设线性可分样本集为 $(x_i, y_i), i = 1, \dots, n; y = \{+1, -1\}$ 是类别标号,分类面方程为:

$$w \cdot x + b = 0 \quad (1)$$

这个平面将两类样本没有错误的分开,并且使得离分类面最近的样本到分类面的距离最大,即分类间隔最大,等价于使 $\|w\|^2$ 最小, w 为分类面的法向量。而要求分类面对所有样本正确分类,约束条件为:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, 2, L, n \quad (2)$$

因此,满足上述条件且使得 $\|w\|^2$ 最小的分类面就是最优分类面。过两类样本中离分类面最近的点且平行于最优分类面的超平面 H_1, H_2 上的训练样本就是式(2)中使等号成立的那些样本叫做支持向量。最优分类面可以表示为如下约束的优化问题,即在式(2)的约束下,求函数

$$\varphi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w) \quad (3)$$

的最小值。为此,可以定义如下的拉格朗日函数:

$$L(w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n a_i \{y_i [(w \cdot x_i) + b] - 1\} \quad (4)$$

(4)式中, $a_i > 0$ 为拉格朗日系数。把原问题转化为如下较简单的对偶问题:

$$\max Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1, j=1}^n a_i a_j y_i y_j (x_i \cdot x_j)$$

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

$$a_i \geq 0, i = 1, \dots, n。$$

1.2 非线性 SVM 上面讨论的是最优和广义线性分类函数,要解决一个特征空间中的最优线性分类问题,我们只需知道这个空间中的内积运算即可。按照广义线性判别函数的思路,要解决一个非线性问题,我们可以设法将它通过非线性变换转换为另一个空间的线性问题,在这个变换空间求最优或最广义分类面。考虑 Mercer 条件:对于任意的对称函数 $K(x, x')$,它是某个特征空间的内积运算的充分必要条件是,对与任意的 $\varphi(x)$ 恒不为0,且 $\int \varphi^2(x) dx < \infty$,有

$\int K(x, y) \varphi(x) \varphi(y) dx dy > 0$, 显然这一条件不难满足^[4]。如果用内积 $K(x, y)$ 代替最优分类面的点积, 就相当于把原特征空间变换到了某一新的特征空间, 此时的支持向量机为:

$$\text{MAX } Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i, x_j)$$

$$\text{s. t. } \sum_{i=1}^n y_i a_i = 0$$

$$0 \leq a_i \leq C \quad i = 1, \dots, n.$$

相应的判别函数也应变为:

$$f(x) = \text{sgn}(\sum_{i=1}^n a_i \cdot y_i K(x_i, x) + b^*).$$

其它的条件不变, 这就是支持向量机。支持向量机的思想可以概括为: 首先通过非线性变换将输入空间变换到一个高维空间, 然后就这个新空间中求取最优线性分类面, 而这种非线性变换是通过定义适当的函数实现的, 这些函数叫做核函数。选择不同的核函数就构成不同的支持向量机, 常用的有以下三类核函数:

$$\text{linear: } K(x, y) = x \cdot y;$$

$$\text{poly: } K(x, y) = [(x \cdot y) + 1]^q;$$

$$\text{rbf: } K(x, y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}.$$

1.3 MATLAB MATLAB 是美国 MathWork 公司推出的一种用于工程计算的高性能程序设计语言。其代码编写过程与数学推导过程的格式很接近。应用主要集中在数值计算、算法开发、数学建模等方面, 以矩阵为运算单元进行计算。MATLAB 作为一种计算工具和科技资源, 可以扩大科学研究的范围、缩短开发周期。该软件的特点语言简洁, 代码灵活, 被称为第四代计算机语言^[2]。其最突出的特点就是提供了更为直观、符合人们思维习惯的代码, 易学易用, 被国际学术界确认为准确、可靠的计算标准软件。

2 支持向量机分类算法的实现

支持向量机算法是在训练样本的特征空间求取能把两类样本没有错误分开的最大间隔超平面, 在数学上表示为一个凸二次规划的问题。也可以说算法求解的主要内容是通过求解二次规划(QP)问题, 这个优化问题的求解是支持向量机算法的核心, 可以说支持向量机的算法就得到了实现。前面所述支持向量机算法可以表示为在式(6)和式(7)的约束下求式(5)取最小值时的拉格朗日乘子 $A = (\partial_1, \partial_2, \dots, \partial_n)^T$, 为训练样本的个数。

$$Q(A) = -A^T I + 1/2 A^T D A \quad (5)$$

$$0 \leq A \leq C \quad (6)$$

$$A^T y = 0 \quad (7)$$

其中: $A = (\partial_1, \partial_2, \dots, \partial_n)^T$ 为 n 元列向量, 是要求的拉格朗日乘子; $D_{ij} = y_i y_j K(x_i, x_j)$ 是一个正定矩阵; $y = (y_1, y_2, \dots, y_n)^T$ 是样本的所属类别, 由 1 或 -1 组成的列向量; x_i 为训练样本。可以看出, 求解支持向量机就是求解上述的一个二次规划问题, 求解后得到拉格朗日乘子 $A = (\partial_1, \partial_2, \dots, \partial_n)^T$, 也就求得了最大间隔超平面。求解这个二次规划问题需要深厚的数学功底数值计算方面的技能, 在主程序语言中实现算法又需要专业的计算机程序设计的知识。在 MATLAB 环境下求解这一问题会变得非常简单, 这得益于 MATLAB 软件强大的优化工具箱, 提供了一个求解二次规划的函数, 可以直接调用。二次规划问题(quadratic programming)的标准形式为:

$$\min f'x + \frac{1}{2} x' H x$$

$$\text{sub. to } Ax \leq b$$

$$Aeqx = beq$$

$$lb \leq x \leq ub$$

其中, H, A, Aeq 为矩阵; f, b, beq, lb, ub, x 为向量, 其它形式的二次规划问题都可转化为标准形式。MATLAB5.x 版中的 qp 函数已被 6.0 版中的函数 quadprog 取代。

函数 quadprog 格式如下:

$$[x, fval] = \text{quadprog}(H, f, A, b, Aeq, beq, lb, ub, x0)$$

其中 $H, f, A, b, Aeq, beq, lb, ub$ 为标准形中的参数; x 为求解得到的最优值, 也就是二次规划的解析解; lb, ub 分别为 x 的下界与上界, 满足不等式约 $lb \leq x \leq ub$; Aeq, beq 满足等约束条件 $Aeq \cdot x = beq$; $x0$ 为设置的初值, 这个值是人为赋予 x 的值, 一般 x 为零; $fval$ 为目标函数最小值, 可以看出, 支持向量机算法是一个标准的二次规划问题; $H = D_{ij} = y_i y_j K(x_i, x_j)$, 根据训练样本数据求出; $f = -1$; 支持向量机算法没形式的的不等式约束条件, 所以 A, b 为空矩阵; $Aeq = A^T, beq = y$, 实现 $A^T y = 0$ 等式约束; $Lb = 0, ub = C$, 实现 $0 \leq A \leq C$ 不等式约束; $x0 = 0$, 赋予 A 的初始值为零。样本数据已知, C 是人工赋予的值。现在支持向量机的求解需要一个公式就可以完成了, 主要 MATLAB 代码如下:

```
function [nsv, alpha, b] = svc(X, Y, ker, C)
%           X           - 训练样本
%           Y           - 训练样本类别
%           ker         - 核函数类型
%           C           - 正则系数
%           nsv         - 支持向量个数
```

```

%          alpha    - 拉格朗日乘子
%          b         - 偏置值
H = zeros(n,n);
for i=1:n
    for j=1:n
        H(i,j) = Y(i)' * Y(j) * svkernel(ker, X(i,:), X
        (j,:));
    end
end
% Dij = yiyjK(xi, xj)
f = -ones(n,1);
lb = zeros(n,1);      % alphas >= 0
ub = C*ones(n,1);     % alphas <= C
x0 = zeros(n,1);      % 赋初值[0 0 0 0]
A = Y', b = 0;        % Ax = b
[alpha lambda how] = qp(H, c, A, b, vlb, vub,
x0, neqctr)           % 调用 qp 函数

```

其中的 `svkernel` 函数为核函数,容易实现,参照核函数的公式编写代码就成。我们调用函数 `svc` 得到拉

参考文献:

- [1] (英)克里斯特安尼. 支持向量机导论[M]. 李国正,王猛,曾华军译. 北京:电子工业出版社,2004.
- [2] 张瑞丰. 精通 MATLAB 6.5[M]. 北京:中国水利水电出版社,2004.
- [3] Vladimir N. Vapnik. 统计学习理论的本质[M]. 张学工译. 北京:清华大学出版社,1999.
- [4] 边肇祺,张学工. 模式识别[M]. 北京:清华大学出版社,1988.

(责任编辑:王瑞斌)

SVM Algorithm Realized in MABLAB

DONG Ting

(Department of Computer and Network Engineering, Yulin College, Yulin 719000, Shaanxi)

Abstract: As a novel - generation - machine - learning - method, Support Vector Machine has caught much attention in recent years, and successfully used in some topics of pattern recognition. In mathematics it presents a quadratic programming. The ways and means to realize SVM in MATLAB are mainly introduced in this paper, and also a simple, efficient, trustful method is offered to the people who study SVM and the people who do the research work.

Key words: SVM; QP; MATLAB

格朗日乘子 $(\alpha_1, \alpha_2, \dots, \alpha_n)^T$ 和偏置 b , 对于任何一个未知样本 x , 用判别函数 $f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b)$ 就可以得到其类别预测值, `sgn` 函数为判别函数,

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \end{cases}.$$

3 结论

MATLAB 软件是数学类应用软件,在数值计算方面尤为突出,被认为是进行高效研究、开发的首选软件工具。加上其语言简单自由,易于学习和掌握,是很多非计算机类专业的科研人员首选开发工具。支持向量机算法在 MATLAB 环境下的实现的核心内容是优化工具箱的应用。基于 MATLAB 环境下支持向量机算法的实现具有方便简单、代码编写和移植快捷、性能可靠、程序运行易于控制等特点。尤其为非计算机专业的科研工作者提供了一种简单、快捷的支持向量机算法研究和应用的技术平台。