Algorithm of Support Vector Machines Based on Statistics Learning Theory

Zhongxiao Hao[1], Xilong Qu[2], Yingchun Liu[2]

1. School of Information Science & Electrical Engineering, Hebei University of Engineering, Handan, China, 056038

2. Department of Computer and Communication, Hunan Institute of Engineering, Xiangtan, China, 411104

1. e-mail h_z_xiao@sina.com, 2. e-mail quxilong@sina.com

Abstract: When the examples are too little, it is difficult to get good effect for machine learning if you use traditional statistics method. So the Statistics Learning Theory developed by Vapnik. is concerned mainly on limited examples. SLT provides us a perfect theory framework for the machine learning problem, at the same time a new general learning algorithm, Support Vector Machines, is developed. Now SVM become a hot area in the machine learning filed. In this article, we mainly introduce the design thought of SVM based on the SLT with the latest research results.

Keywords: statistics learning theory，support vector machines，pattern classification

基于统计学习理论的支持向量机算法

郝忠孝[1]，屈喜龙[2]，刘迎春[2]

1. 河北工程大学信息与电气工程学院，邯郸，中国，056038

2. 湖南工程学院计算机与通信学院，湘潭，中国，411104

1. e-mail h_z_xiao@sina.com., 2. e-mail quxilong@126.com

【摘要】传统的统计方法当样本数目有限时难以取得理想的效果，由 Vapnik 等人提出的统计学习理论着重研究在小样本情况下的统计学习方法。统计学习理论为机器学习问题建立了比较完善的理论框架，同时发展出一种通用的学习方法－支持向量机。目前，SVM 已成为机器学习领域中的研究热点。本文结合最新的 SVM 研究成果从统计理论出发来阐述支持向量机的设计思想。

【关键词】统计学习理论；支持向量机；模式分类

1 引言

支持向量机(Support Vector Machines，SVM)是 20 世纪 90 年代由 Vapnik 等人提出的基于统计学习理论(Statistic Learning Theory，SLT)的一种新的学习机[1]，它是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度，Accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的推广能力(Generalization Ability，是指将学习机器对未来输出进行正确预测的能力)。支持向量机方法的几个主要优点有：专门针对有限样本情况的；算法最终将转化成为一个凸二次规划问题，从理论上说，得到的将是全局最优点；有较好的推广能力和避免了维数灾难。本文主要从统计学习理论来阐述支持向量机的算法思想，主要针对的是模式分类问题。

项目支持：湖南省教育厅资助科研项目(08A009)和(08B015)、湖南省重点学科建设项目资助、湖南省自然科学基金资助项目(01JJY2157)；湖南工程学院党建与思想政治教育研究项目(D08-05).

2 统计学习基础

本机器学习的目的是根据给定的训练样本求对某系统输入输出关系的估计，使它能够对新的输入做出尽可能准确的预测，一般表示为：输出 y 与输入 x 之间存在一定的未知关系，即遵循某一未知的联合概率分布 $P(x,y)$，机器学习问题就是根据 l 个独立同分布样本在假设集 $\mathrm{F}=\{f(x,w):f(x)=(w\cdot x)+b\}$ 中求一个最优的 $f(x,w_0)$ 使期望风险最小。作为对期望风险的近似。但是在有限样本情况下，存在经验风险最小未必意味着期望风险最小，并且学习机器的复杂性不但应与所研究的系统有关，也要和有限数目的样本相适应。而由 Vapnik 等人发展起来的 VC

978-1-935068-01-3 © 2009 SciRes.

理论与结构风险最小化原则比较理想的解决了这个问题。

为了研究学习过程一致收敛的速度和推广性，统计学习理论定义了一系列有关假设集学习性能的指标，其中最重要的是 VC 维。模式识别方法中 VC 维的直观定义是：对一个指示函数集，如果存在 h 个样本能够被假设集中的函数按所有可能的 2^h 种形式分开，则称假设集能够把 个样本打散；假设集的 VC 维就是它能打散的最大样本数目。VC 维反映了假设集的学习能力，VC 维越大则学习机器越复杂。

实现 SRM 原则可以有两种思路，一是在每个子集中求最小经验风险，然后选择使最小经验风险和置信范围之和最小的子集。显然这种方法比较费时，当子集数目很大甚至是无穷时不可行。因此有第二种思路，即设计函数集的某种结构使每个子集中都能取得最小的经验风险(如使训练误差为 0)，然后只需选择选择适当的子集使置信范围最小，则这个子集中使经验风险最小的函数就是最优函数，支持向量机方法实际上就是这种思想的具体实现。

3 支持向量机

支持向量机的基本思想是对于非线性可分样本，将其输入样本通过非线性变换映射到另一个高维空间 Z 中，在变换后的空间中构造一个最优的分类超平面，使其在保证分类精度(满足经验风险)的同时最大化超平面两侧的空白区域(即最大化置信范围)，也即是 H1 与 H2 间的几何间隔。这使得分类的结果不但在训练集上得到优化，而且在整个样本集上的风险也有上界。

对于线性可分两类模式识别问题中，给定训练数据**错误！未找到引用源。**，支持向量机就是指根据结构风险最小化原则确定的分类假设(也称为决策函数) $f(x)=(w\cdot x)+b$ 输出 $y=\mathrm{sgn}(f(x))$,sgn() 是符号函数，决策函数 $f(x)$ 是特征空间某点 x 到超平面的距离的一种代数度量。下面给出超平面间隔(*margin*)的定义，记：

$$\begin{aligned} &margin = r_+ + r_- \\ &r_+ = \min\left\{\frac{w^T x_i + b}{\|w\|} \mid i \in \{1,2,\cdots,m \mid y_i = +1\}\right\} \\ &r_- = \min\left\{\frac{w^T x_i + b}{\|w\|} \mid i \in \{1,2,\cdots,n \mid y_i = -1\}\right\} \end{aligned} \tag{1}$$

当存在点使得 $w^T x_i + b = \pm 1$ 时，$r_+ = r_- = 1/\|w\|$，从而超平面的 $margin = 2/\|w\|$。对于给定的样本集 X，如果存在分离超平面 $(w\cdot x)+b=0$ 使其间隔最大，则称该超平面为最优分类超平面。

支持向量机的基本思想，就是在保证精度的同时最大化超平面两个的空白区域(即几何间隔)。在样本集线性可分的情况下，我们可以得到(最大化 $2/\|w\|$ 也即是最小化 $\|w\|^2/2$)

$$\begin{aligned} &\min_{\omega,b} \quad \tau(w) = \frac{1}{2}\|w\|^2 \\ &s.t. \quad y_i((w\cdot x_i)+b) \ge 1, i = 1,2,\cdots,l \end{aligned} \tag{2}$$

式（2）是最大间隔原则，其中的 b 是阈值，w 是权重向量。

从上两式可以看出，支持向量机的核心思想是将分类问题转化为最优化问题，通过最大化间隔来构造最优超平面，使分类间隔最大就是对推广能力的控制。统计学习理论指出，在 N 维空间中，设样本分布在一个半径为 ζ 的超球范围内，则满足条件 $\|w\| \le \Lambda$ 的正则超平面构成的假设集 $\mathcal{F}$ 的 VC 维满足下面的不等式

$$h \le \min([\zeta^2 \Lambda^2], N) + 1 \tag{3}$$

因此使 $\|w\|^2$ 最小就是使 VC 维的上界最小，从而实现 SRM 准则中对假设集复杂性的选择（就是使得**错误！未找到引用源。**式中的 $\Psi(h/l)$ 最小）。

式（2）的学习机是针对线性可分的情况的，由于线性可分时没有错分样本，因此不考虑经验风险(此时经验风险为 0)。而对于线性不可分问题，通过引入松弛变量 ξ_i，可得“软化”了的约束条件：

$$\begin{aligned} &y_i((w\cdot x_i)+b) \ge 1-\xi_i, \\ &\xi_i = \max(0, \frac{1}{\|w\|} - y_i((w\cdot x_i)+b)), i = 1,\cdots,l \end{aligned} \tag{4}$$

同时在最大化间隔中对其进行惩罚，这样，原始问题（3）可以改为

$$\begin{aligned} &\min_{\omega,b,\xi} \quad \tau(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i((w\cdot x_i)+b) \ge 1-\xi_i, i = 1,\cdots,l \\ &\qquad \xi_i \ge 0, i = 1,\cdots,l \end{aligned} \tag{5}$$

其中 $C>0$ 是一个惩罚参数，当 w 是不可行点时，在 w 处 $C\xi_i$ 是很大的正数，它的存在是对点脱离可行域的一种惩罚，其作用是在最小化过程中迫使迭代点靠近可行域。它控制对错分样本惩罚的程度，使得样本偏差与机器推广能力之间寻求一个平衡。问题（5）的 Wolfe 对偶问题是：

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^{l} \alpha_i \\ s.t. \quad & \sum_{i=1}^{l} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C, \quad i = 1, \cdots, l \end{aligned} \tag{4}$$

α_i 是拉格朗日乘子。最后求得：

$$w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i, \quad b^*_{0<\alpha_j^*<C} = y_j - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j)$$

如果不存在 $0 < \alpha_j^* < C$，则 b 应该在一个区间上。

显然由式（1）构造的分割超平面，仅仅依赖于那些对应于 α_i^* 不为 0 的训练点 (x_i, y_i)，而与对应于 α_i^* 为 0 的训练点无关。这些点称为支持向量(support vectors)，它们是训练集合中仅有的有效数据点。如果问题**错误!未找到引用源。**的一个解 $\alpha = (\alpha_1, \cdots, \alpha_l)^T$，它与输入 x_i 对应的 $\alpha_i > 0$，则称输入 x_i 为支持向量。如图 2 中，H_1、H_2 上的训练样本点就是支持向量。

对非线性问题，可以通过非线性变换转化为某个高维空间中的线性问题，在变换空间求最优分类面。注意到在上面的对偶问题中，不论是寻优目标函数（6）还是分类函数（1）都只涉及训练样本之间的内积运算。根据泛函的有关理论，只要一种核函数 $K(x_i, x_j)$ 满足 Mercer 条件，它就对应某一变换空间中的内积。因此，对于非线性的问题，只需要引入核函数即可，对于结构风险最小化原则的分析和线性可分问题没有太多的区别。这一特点提供了解决算法可能导致的“维数灾难”问题的方法：在构造决策函数时，传统方法是对输入空间的样本作非线性变换，然后在特征空间中求解；而支持向量机是先在输入空间比较向量(例如求内积或是某种距离)，对结果再作非线性变换。这样，大的工作量将在输入空间而不是在高维特征空间中完成。同时，通过把原问题转化为对偶问题，计算的复杂度不再取决于空间维数，而是取决于样本数，尤其是样本中的支持向量数。

在邓乃扬的文献中[3]提出了一个定理，假设概率分布 P_x 满足：

$$P_x\{x : \| x \| \le \zeta\} = 1 \tag{7}$$

若考虑线性决策函数 $f(x) = \mathrm{sgn}((w \cdot x) + b)$，则的任给 $\delta \in (0,1]$，相对于 0-1 损失函数，期望风险 $R(w)$ 至少以 $1-\delta$ 的概率满足

$$R(w) \le \frac{2}{l}([d_{eff} \log_2(\frac{8el}{d_{eff}}) \log_2(32l)] + \log 2(\frac{(16 + \log_2 l)l}{8}))$$

这里 $d_{eff} \le 2l$。这个定理给出期望风险 $R(w)$ 的至少以 $1-\delta$ 概率成立的上界，显然这个上界是 d_{eff} 的单调增函数，因此在选择决策函数时，应使 d_{eff} 尽可能小。考虑问题（6）目标函数的一变形：

$$\min_{\omega, b, \xi} \quad \tau(w) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i^2 \tag{8}$$

其中 $C > 0$。这个问题与问题（6）不完全一致，但本质是相同的，只是对应不同的损失函数。考察式（7）和 d_{eff} 的关系，可以发现，d_{eff} 由两项构成，分别对应于式（8）中目标函数的第一项和第二项中的 $\sum_{i=1}^{l} \xi_i^2$，目标函数中的参数 C 则是调节以上两个因素 $\| w \|^2$ 和 $\sum_{i=1}^{l} \xi_i^2$ 的因子。参数 C 越大，惩罚越大，对错分样本的约束程度就越大，得到分类面的间隔就越小；随着 C 的降低，支持向量机忽略更多的样本，得到较大边缘间隔的分类面。所以，这里的 $\sum_{i=1}^{l} \xi_i^2$ 体现了经验风险，而 $\| w \|$ 则体现了表达能力。所以惩罚参数 C 实质上是对经验风险和表达能力如何匹配的一个裁决。从以上的分析可以看出，支持向量机是以结构风险最小化原则为指导而设计出的一种学习机器，具有良好的推广能力。

4 结束语

由于支持向量机不需要先验知识和大样本，它在文本分类、图像分类、生物序列分析和生物挖掘、手写字符识别、遥感图像分析中得到比较大的应用。特别是贝尔实验室对美国邮政手写数字库进行的实验。人工识别平均错误率是 5.9%，决策树方法错误率是 16.2%，神经网络方法错误率是 5.9%，专门针对该问题设计的五层神经网络方法错误率是 5.1%，而采用 SVM 方法的错误率在 4.1%。实验证明 SVM 相对于传统方法有明显的优势。相对于 SVM 的理论研究来说，其应用研究还没有得到大规模的开展。

当前的 SVM 的研究，一是通过对其本身理论的研究，提出进一步完善的措施，其中包括有多类识别问题和快速训练算法[4][5]等。另外就是探索新的应用领域，SVM 本质上是一种非线性数据处理工具，在数字信号处理、图像识别、智能控制等领域有巨大的应用潜力，在这方面现在已经有了一些成果。

致谢

感谢电子科技大学罗瑜博士对本文提供的宝贵意见，也感谢参考文献中所列的每一位作者。

References (参考文献)

[1] Thorsten Joachims. Making Large-Scale SVM Learning Practical. [J] In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999. 169−184.

[2] Fan, R.-E., Chen, P.-H., & Lin, C.-J.. Working set selection using second order information for training SVM[J]. Journal of Machine Learning Research, 2005, 6, 1889−1918.

[3] E. Osuna, R. Freund, F. Girosi. An improved training algorithm for support vector machines[J]. Proc.IEEE Workshop on Neural Networks and Signal Processing, Piscataway: IEEE Press, 1997, pp. 276−285.

[4] C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998.

[5] Zhang Ling. The relationship between kernel functions based SVM and three-layer feedforward neural networks[J]. Chinese J Computer, 2002, 25(7): 1−5.

[6] Zhang Ling, Zhang Bo. Relationship between support vector set and kernel functions in SVM[J]. J Comput Sci & Technol, 2002, 17(5): 549−555.

[7] Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design[J]. Neural Computation, 2001, 13: 637−649.

[8] Keerthi S S, Gilbert E G. Convergence of a generalized SMO algorithm for SVM classifier design[J]. Machine Learning, 2002, 46(1): 351−360.

[9] Platt J C. Fast training of support vector machines using sequential minimal optimization[A]. SCHLKOPH B, et al eds. Advances in Kernel Method-Support Vector Learning[C]. Cambridge, MA: MIT Press, 1999. 185−208.

[10] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines (EB/OL), http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[11] Laskov P. Feasible direction decomposition algorithms for training support vector machines[J]. Machine Learning, 2002, 46(1): 315−349.

[12] Hsu C-W, LIN C-J. A simple decomposition method for support vector machines[J]. Machine Learning, 2002, 46(1): 291−314.

[13] Lin C-J. On the convergence of the decomposition method for support vector machines[J]. IEEE Trans on Neural Networks, 2001, 12(6): 1288−1298.

[14] Ayat N E, CHERIET M, REMAKI L, et al. KMOD-a new support vector machine kernel with moderate decreasing for pattern recognition, application to digit image recognition[A]. Proceedings of 6th Int Conf on Document Analysis and Recognition[C]. Seattle, USA: IEEE, 2001. 1215−1211.

[15] Amari S-I, WU S. An information-geometrical method for improving the performance of support vector machine classifier[A]. Proceedings of 9th Int conf on Artificial neural networks[C]. Edinburgh, UK: IEEE, 1999. 85−90.

[16] Chapelle O, Vapnik V, Bacsquest O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1): 131−159.

[17] Baesens B, Viaene S, Gestel T V, et al. An empirical assessment for kernel type performance for least squares support vector machine classifiers[A]. Proceedings of 4th Int Conf on Knowledge-based Intelligent Engineering Systems and Allied Technologies[C]. Brighton, UK: IEEE, 2000. 313−316.

[18] Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2): 415−425.

[19] Sebald D J, BUCHLEW J A. Support vector machines and the multiple hypothesis test problem[J]. IEEE Trans on Signal Processing, 2001, 49(11): 2865−2872.