基于支持向量机的分段线性学习方法

杨 强 吴中福 王 茜
(重庆大学计算机学院 重庆400044)

A Subsection Learning Algorithm Based on Support Vector Machines

YANG Qiang WU Zhong-Fu WANG Qian
(Academy of Computer, Chongqing University, Chongqing 400044)

Abstract In this paper, we discuss drawback of traditional subsection learning algorithm in pattern recognition and exiting support vector machines (including kernel functions), the necessity of using subsection learning algorithm based on support vector machines as well as. In turn, a subsection learning algorithm based on support vector machines, is proposed in this paper.

Keywords Support vector, Subsection linear, Small samples, Statistical learning

1. 引言

包括感知器、神经网络等在内的学习方法都是基于经验风险最小(ERM)原则的[1],而在实际的基于小样本的学习系统中,这些学习方法在经验风险最小的情况下并不能保证期望风险最小化。对于线性不可分情况不能给出是否分段线性可分的可靠信息。如果简单地引入非线性变换,则容易导致过学习现象。这显然不是我们所希望的。

上世纪九十年代中期,由于支持向量机的诞生[2,3],使得我们对线性可分系统,能够在结构风险最小化(SRM)原则下,通过构造最优超平面来最小化期望风险。对于线性不可分情况目前也有添加松弛变量和引入核函数[4,5]两种方法。但是仅有这两种方法显然是不完善的,因为现实生活中存在大量分段线性可分系统。而已有的这两种方法并没有给出对此类问题的解决方法,如果简单地套用已有的方法将导致正确分辨率较低或期望风险较大的后果,因此本文提出基于支持向量的分段线性学习方法,以期在综合考虑经验风险和结构风险的情况下,给出对分段线性可分系统最佳的解决方法。

2. 现有的基于支持向量机的学习方法及其局限性

上世纪九十年代中期,由 Vladimir N. Vapnik 等人在 VC 维和结构风险最小化原则的基础上创造了支持向量机(SVM)的学习方法。由于 SVM 具有非渐进收敛的特性,使得以 SVM 为基础的统计学习理论成为研究模式识别、回归分析和密度估计的非常有效的工具,近年来成为一个非常活跃的研究领域。

对于线性可分系统,利用构造最优超平面的技术,将模式识别转换为二次规划问题。为了对本文将要提出的分段线性学习算法有清晰的认识,我们首先讨论这个二次规划问题,为此先引入 Δ-间隔分类超平面概念。

Δ-间隔分类超平面:一个超平面 $(\vec{w}^*\cdot\vec{x})-b=0$, $\|\vec{w}^*\|=1$,如果它以如下形式将向量 $\vec{x}$ 分类:

$$y=\begin{cases}1 & (\vec{w}^*\cdot\vec{x})-b\geqslant\Delta\\ -1 & (\vec{w}^*\cdot\vec{x})-b\leqslant-\Delta\end{cases}$$

则我们称之为 Δ-间隔分类超平面。令 $\Delta=1/\|\vec{w}^*\|$,对训练数据 $(\vec{x}_1,y_1),\cdots,(\vec{x}_l,y_l)$, $\vec{x}\in R^n$, $y\in\{+1,-1\}$,得到正规化形式:

$$y_i[(\vec{w}^*\cdot\vec{x})-b]\geqslant1,\ i=1,\cdots,l$$

且有如下的定理与推论成立:

定理1 向量 $\vec{x}\in X$ 属于一个半径为 R 的球中,那么 Δ-间隔分类超平面集合的 VC 维 h 以下面的不等式为界:

$$h\leqslant\min([R^2/\Delta^2],V)+1$$

其中,V 是空间的维数。

推论 以概率 $1-\eta$ 我们可以断定,测试样本不能被 Δ-间隔分类超平面正确分类的概率有下面的界:

$$P_{error}\leqslant\frac{m}{l}+\frac{\Psi}{2}\left[1+\sqrt{1+\frac{4m}{l\Psi}}\right],$$

其中,$\Psi=4(h(\ln\frac{2l}{h}+1)-\ln\frac{\eta}{4})/l$

m 是没有被 Δ-间隔分类超平面正确分类的训练样本数目。

因此,对于线性可分系统,我们可以将构造最优超平面问题转换为二次规划问题:最小化泛函:

$$\Phi(\vec{w})=\frac{1}{2}(\vec{w}\cdot\vec{w})$$

约束条件为不等式类型:$y_i[(\vec{x}_i\cdot\vec{w})-b]\geqslant1$, $i=1,2,\cdots,l$。其中只有一部分所谓的支持向量才具有非零的系数 α_i^0 使得这个不等式的等式成立。通过上面的二次规划求解,可得:

$$\vec{w}_0=\sum_{\text{支持向量}}y_i\alpha_i^0\vec{x}_i,\ \alpha_i^0\geqslant0$$

由 Kühn-Tucker 条件可知,α_i^0 为最优化以下泛函的解:

$$W(\vec{\alpha})=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_jy_iy_j(\vec{x}_i\cdot\vec{x}_j)$$

约束条件为:$\alpha_i\geqslant0$, $i=1,\cdots,l$ 和 $\sum_{i=0}^{l}\alpha_iy_i=0$。

设 $\vec{\alpha}_0=(\alpha_1^0,\cdots,\alpha_l^0)$ 为这个二次优化问题的解,那么与最优超平面对应的向量 $\vec{w}_0$ 的模等于:

$$\|\vec{w}_0\|^2=2W(\vec{\alpha}_0)=\sum_{\text{支持向量}}\alpha_i^0\alpha_j^0(\vec{x}_i\cdot\vec{x}_j)y_iy_j$$

基于最优超平面的分类规则为下面的指示函数:

$$f(\vec{x})=\text{sgn}\left(\sum_{\text{支持向量}}y_i\alpha_i^0(\vec{x}_i\cdot\vec{x})-b_0\right),$$

其中 $\vec{x}_i$ 是支持向量,α_i^0 是对应的拉格朗日系数,b_0 是常数(阈值):

$$b_0=\frac{1}{2}[(\vec{w}_0\cdot x^*(1))+(\vec{w}_0\cdot x^*(-1))]$$

其中，$\vec{x}^*(1)$为属于第一类的任意一个支持向量，$\vec{x}^*(-1)$为属于第二类的任意一个支持向量。由此我们对线性可分系统给予了圆满的解决。

对于线性不可分系统，已有两种基于 SVM 的解决方法，其中一种为加入松弛变量 $\xi_i\geqslant 0$和最小化泛函：

$$F_\sigma(\xi)=\sum_{i=1}^{l}\xi_i^\sigma$$

其中参数 $\sigma>0$。约束条件为 $y_i[(\vec{x}_i\cdot\vec{w})-b]\geqslant 1-\xi_i, i=1,2,\cdots,l$，和另一个条件$(\vec{w}\cdot\vec{w})\leqslant\Delta^{-2}$。

对足够小的 $\sigma>0$，这个优化问题的解定义了这样一个超平面，它的参数属于上述第二个约束定义的子集

$$S_n=\{(\vec{w}\cdot\vec{x})-b:(w\cdot w)\leqslant\Delta^{-2}\}$$

(即属于由常数 $c_n=\Delta^{-2}$决定的结构的元素)，在这种情况下这个超平面使得训练错误数最小。在此基础上提出了构造 Δ-间隔分类超平面和构造软间隔分类超平面两种方法。对于 Δ-间隔分类超平面方法在下一节将详细论述。由于本文并不采用构造软间隔分类超平面方法，所以在此不再熬述。

但是此类方法中，对于 Δ-间隔分类超平面和构造软间隔分类超平面两种方法中各自的常数 $c_n=\Delta^{-2}$和系数 C 的确定是一个较难解决的问题。由此也就增加了学习系统的难度和降低了学习系统的精确性和可靠性。

另一种方法是对线性不可分系统用核函数将输入空间映射到特征空间求解。虽然对于本质非线性系统，用核函数方法当然是很好的选择，但是由于核函数的引入将导致以下几个问题。首先，要用核函数进行模式识别就得在特征空间中对

$$\Phi(R,\vec{w}_0,l)=R^2\|\vec{w}_0\|^2/l$$

实现最小化。其中，

$$R^2=R^2(K)=\min_{\vec{a}}\max_{\vec{x}_i}\lfloor K(\vec{x}_i,\vec{x}_j)+K(\vec{a},\vec{a})-2K(\vec{x}_i,\vec{a})\rfloor$$

$\vec{a}$是包含所有样本的最小超球的中心。这比在输入空间中最大化

$$W(\alpha,C^*)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2C^*}\sum_{i,j=1}^{l}\alpha_i\alpha_jy_iy_j(\vec{x}_i\cdot\vec{x}_j)-\frac{C^*}{2\Delta^2}$$

(这个等式将在第三节中进行严密的推导)要复杂得多。其次，由于核函数的引入会导致 VC 维的增加，这就会增加学习系统的结构风险。即使核函数采用内积回旋的形式，使得 VC 维的上界等于学习样本的个数 l。但是对于学习样本数量太大的系统而言，学习系统的 VC 维将增大，导致我们对学习系统的结构风险难以估计。再其次，从计算 R^2的式子可知，用一般的微机系统，由于内存的限制是难于用核函数方法对4000个以上的样本进行有效计算的。最后，为了让非线性系统在特征空间中线性可分，就得确定好的核函数，这对于提高系统的学习性能是至关重要的。但是，只有基于较强的先练知识，才能确定好的核函数。这几方面都无疑增加了构造好的基于核函数的学习系统的复杂性。

3. 基于支持向量机的分段线性学习方法

由前面的分析可知，为了提高模式识别系统的可靠性，对于在输入空间中不能用超平面进行正确分类的系统，以及通过引入核函数，但由于样本分布过于复杂，或者由于引入的核函数不太合适，在特征空间中仍然不能用超平面正确分类的情况，就得采用新的模式识别方法。因此本文提出利用贪婪思想的基于 SRM 原则的支持向量分段线性学习方法。(算法中将首先用构造 Δ-间隔分类超平面的方法进行初步学习，根据所确定的支持向量和错分区域确定分段线性的节点，注意：不能用线性可分系统的方法，因为线性可分系统的方法在线性不可分系统的学习中将是不收敛的，由于得不出最优解，将导致死循环。)

按照前面在输入空间中，对于用超平面不能正确划分的情况，首先采用引入非负松弛变量 $\xi_i\geqslant 0$和最小化泛函：

$$F_\sigma(\xi)=\sum_{i=1}^{l}\xi_i^\sigma,$$

其中参数 $\sigma>0$。为了计算的原因，令 $\sigma=1$。

约束条件为 $y_i[(\vec{x}_i\cdot\vec{w})-b]\geqslant 1-\xi_i, i=1,2,\cdots,l$，和另一个条件$(\vec{w}\cdot\vec{w})\leqslant\Delta^{-2}$。

只是这里确定的常数 $c_n=\Delta^{-2}$比文[2]中的方法要随意得多，我们可以取比采用文[2]的方法更小的值，以降低学习系统的 VC 维，从而减少学习系统的结构风险。利用 Kühn-Tucker 定理可知，$F(\xi)$在以下的拉格朗日泛函：

$$L=\sum_{i=1}^{l}\xi_i-\frac{C^*}{2}(\Delta^{-2}-\vec{w}\cdot\vec{w})-\sum\alpha_i\{[(\vec{w}\cdot\vec{x}_i)+b]y_i-1+\xi_i\}-\sum_{i=1}^{l}r_i\xi_i$$

的鞍点上得到。其中，$\alpha_i\geqslant 0, r_i\geqslant 0, i=1,\cdots,l, C^*\geqslant 0$，根据

$$\frac{\partial L}{\partial\vec{w}}=0,\ \frac{\partial L}{\partial b}=0,\ \frac{\partial L}{\partial\xi_i}=0,$$

得到

$$\vec{w}=\frac{1}{C^*}\sum_{i=1}^{l}\alpha_iy_i\vec{x}_i,\quad \sum_{i=1}^{l}y_i\alpha_i=0, 0\leqslant\alpha_i\leqslant 1$$

将 $\vec{w}=\frac{1}{C^*}\sum_{i=1}^{l}\alpha_iy_i\vec{x}_i$ 代入拉格朗日泛函得到参数 $\alpha_i, i=1,\cdots,l$ 和 C^*是下面的凸优化问题的解：

最大化泛函

$$W(\alpha,C^*)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2C^*}\sum_{i,j=1}^{l}\alpha_i\alpha_jy_iy_j(\vec{x}_i\cdot\vec{x}_j)-\frac{C^*}{2\Delta^2}$$

约束条件为$\sum_{i=1}^{l}y_i\alpha_i=0, C^*\geqslant 0, 0\leqslant\alpha_i\leqslant 1, i=1,\cdots,l$。

本文的算法是基于上述的 Δ-间隔分类超平面方法思想，初略给出较小的 Δ^{-2}，即使在开始的学习中有较多的错误分类也无关紧要(因为在算法的循环学习中将给予改进)。然后再按照贪婪思想进行迭代，以使错误分类数少到满意为止。这样我们就在 SRM 原则下，将经验风险限制在容许的范围之内。

因此，本文提出的分段线性学习算法分为以下步骤：

(1)确定常数 $c_n=\Delta^{-2}$(可以取比较小的值)和允许错分的样本数 m 和 $F(\xi)=\sum_{i=1}^{l}\xi_i$ 的上界。令 $k=1$。

(2)对每个划分的子区域 n(当 $k=1$时，将所有学习样本看作一个区域)，用前面讲述的 Δ-间隔分类超平面方法求解 $f_k^n(\vec{x})$、$\vec{w}_k^n$ 和 b_k^n。如果 $C_{k,n}^*$和 $\alpha_{i,k}^n$是个子区域凸优化问题的解，那么，

$$f_k^n(\vec{x})=\mathrm{sgn}\left(\frac{1}{C_{k,n}^*}\sum_{\text{支持向量}}y_i\alpha_{i,k}^n(\vec{x}_i\cdot\vec{x})-b_k^n\right),$$

$$\vec{w}_k^n=\frac{1}{C_{k,n}^*}\sum_{\text{支持向量}}\alpha_{i,k}^ny_i\vec{x}_i,$$

$$b_k^n=\frac{1}{2}[(\vec{w}_k^n\cdot x^*(1))+(\vec{w}_k^n\cdot x^*(-1))]$$

其中，$\vec{x}^*(1)$为属于第 k 次学习中、第 n 子区域、第一类的任

意一个支持向量，$\vec{x}^*(-1)$为相应的属于第二类的任意一个支持向量。

(3)检查 $m_k\leqslant m$ 和 $F_k(\xi)\leqslant F(\xi)$是否成立。其中，m_k 为第 k 次学习中错误分类的训练样本数目。$F(\xi)$为第 k 学习中得到的最小值。

(4)如果步(3)中的两个不等式均成立，则停止学习过程。如果任意一个不等式不成立，则标记出分类错误的点，寻找出各个子区域的紧互对，并对错误分类的点比较多的子区域计算出分段节点。对划分后的各个子区域 $k=k+1$，转到(2)。

由上面的算法可知，利用该算法可以很好地控制经验风险，下面将证明该算法也很好地控制了结构风险。从而也就很好地控制了期望风险限。

证明：根据定理1：

$$h\leqslant \min([\frac{R^2}{\Delta^2}],V)+1,$$

成立。对任意 $R_n\geqslant 0, n=1,\cdots,N_k$，又有不等式：

$$\sum_{n=1}^{N_k} R_n^2\leqslant(\sum_{n=1}^{N_k} R_n)^2\approx R^2$$

成立。这里 $R_n\geqslant 0, n=1,\cdots,N_k$，表示各个子区域的半径。$N_k$ 为第 k 次学习中子区域的数目。又因为，对所有的子区域 $\Delta\leqslant\Delta_n$，所以 $h\geqslant\sum_{n=1}^{N_k} h_n$，因此这个算法的期望风险得到了有效控制。

从算法及其证明可知，这种算法相对于 Δ-间隔分类超平面方法而言，可以提高识别正确率。既降低经验风险，相对于核函数方法而言，可以降低结构风险又能降低模式识别系统的复杂度，也避免了核函数方法对多样本系统不能有效处理的问题。所以采用本文提出的方法，我们可以比较好地限制期望风险，从而构造出好的学习系统。

结论 本文结合支持向量和邻域法思想提出了基于支持向量的分段线性的学习算法。这种算法克服了传统的分段线性的模式识别方法由于不能控制学习系统的结构复杂性，而可能出现的过学习现象，也避免了现有的基于支持向量的学习算法不能较好控制经验风险或较难确定好的核函数的缺点。另外，即使对于本质非线性系统也有助于降低对所引入的核函数的要求和降低学习系统的结构复杂性，并通过理论证明这种算法是正确的。

本文提出的算法，在构造性控制子区域的数量和合理性方面还有待于进一步改进，以适应复杂样本分布的情况。

参考文献

1 边肇祺，张学工．模式识别．清华大学出版社，2000

2 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1999

3 Vapnik V N. An Overview of Statistical Learning Theory. IEEE Translations on Neural Networks, 1999, 10(5): 988～999

4 Scholkopf B. Input Space Versus Feature Space in Kernel-Based Methods. IEEE Translations on Neural Networks, 1999, 10(5): 1000～1016

5 Muller K-R. An Introduction to Kernel-Based Learning Algorithms. IEEE Translations on Neural Networks, 2001, 12(2): 181～201

(上接第98页)

4 Lempel R, Moran S. The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect. In: Proc. of the 9th Intl. World Wide Web Conf. 2000

5 Gallager R G. Discrete Stochastic Processes, Kluwer Academic Publishers, 1996

6 Borodin A, Roberts G O, Rosenthal J S, Tsaparas P. Finding Authorities and Hubs From Link Structure on the World Wide Web. In: Proc. of the 9th Intl. World Wide Web Conf. 2000

7 Marchiori M. The Quest for Correct Information on the Web: Hyper Search Engines. In: The Sixth Intl. WWW Conf. (WWW97), Santa Clara, USA, 1997

8 Mukherjea S, Foley J D. Showing the Context of Nodes in the World Wide Web. In: Proc. of ACM CHI'95 Conf. on Human Factors in Computing Systems, volum 2 of short papers: Web Browsing, 1995

9 Pirolli P, Pitkow J, Rao R. Silk from a Sow's Ear: Extracting Usable Structures from the Web. In: Proc. of 1996 Conf. on Human Factors in Computing Systems (CHI96), Vancouver, British Columbia, Canada, 1996

10 Brin S, Page L. The Anatomy of Large-Scale Hypertextual Web Search Engine. In: Proc. of 7th Intl. World Wide Web Conf. 1998

11 Weiss R, Velez B, Sheldon M A. Hypursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. In: Proc. of the 7th ACM Conf. on Hypertext, New York, 1996

12 Chakrabarti S, et al. Experiments in Topic Distillation. In: Proc. of 8th Intl. World Wide Web Conf. 1999

13 Dean J, Henzinger M R. Finding Related Pages in the World Wide Web. In: Proc. of the 8th Intl. World Wide Web Conf. Toronto, Canada, Elsevier Science, 1999. 5

14 Bharat K, et al. The Connectivity Server: fast access to linkage information on the Web. In: Proc. of 7th Intl. World Wide Web Conf. 1998

15 Chakrabarti S, Dom B, Indyk P. Enhanced hypertext categorization using hyperlinks. In: Proc. of the 1998 ACM Intl. Conf. on Management of Data (SIGMOD'98), 1998

16 Bharat K, Henzinger M R. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proc. of the 21st Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1998

17 Zhang D, Dong Y. An Efficient Algorithm to Rank Web Resources. In: Proc. of 9th Intl. World Wide Web Conf. Amsterdam, 2000

18 Rafiei D. What is this Page Known for? Computing Web Page Reputations. In: Proc. of 9th Intl. World Wide Web Conf. Amsterdam, 2000

19 Dwork C, Kumar R, Naor M, Sivakumar D. Rank Aggregation Methods for the Web. In: Proc. of 10th Intl. World Wide Web Conf. Hong Kong, 2001

20 Chakrabarti S, Joshi M, Tawde V. Enhanced Topic Distillation using Text, Markup tags, and Hyperlinks, ACM SIGIR 2001, New Orleans, 2001

21 Bharat K. When Experts Agree: Using Non-affiliated Experts to Rank Popular Topics. In: Proc. of 10th Intl. World Wide Web Conf. Hong Kong, 2001

22 Chakrabarti S, et al. Mining the Link Structure of the World Wide Web, IEEE Computer, 1999. 8

23 Mukherjea S. WTMS: A System for Collecting and Analyzing Topic-Specific Web Information. Computer Networks, 2000, 33

24 Spertus E. ParaSite: Mining Structural Information on the Web. In: Proc. of 6th Intl. World Wide Web Conf. Santa Clara, CA, 1997

25 chakrabarti S, Dom B, Raghavan P, Rajagopalan S. Automatic Resource Compilation by analyzing Hyperlink Structure and Associated Text. In: Proc. of the 7th Intl. World Wide Web Conf. 1998