文章编号:1000 - 582X(2004)05 - 0041 - 04

基于正反馈的支持向量机*

杨 强,吴中福,余 平,钟 将
(重庆大学 计算机学院 重庆 400030)

摘 要:在分析现有的基于高斯核的支持向量机(包括基于 K - 邻域法的支持向量机)的优缺点的基础上,通过对支持向量机之所以能够描述数据集的分布特征的本质进行分析,突破目前在构造支持向量机中存在的"所有支持向量与样本之间的在特征空间中的内积所对应的核函数参数一定要相等"的这一苛刻要求,提出了用于模式识别的基于正反馈的支持向量机。给出了基于正反馈的支持向量机的算法。通过对人工数据和现实数据的仿真实验,表明基于正反馈的支持向量机在推广性能方面明显优于现有的支持向量机。

关键词:径向基神经网络;K - 邻域法;核函数;支持向量机
中图分类号:TP181 **文献标识码**:A

径向基神经网络在神经网络的发展和应用中,一直处于非常重要的地位。在径向基神经网络中,它的隐单元的个数和各个隐单元所对应的权值,只能通过利用经验风险最小化原则来确定。众所周知,由于学习过程中常常存在欠学习或过学习现象,使得经验风险与神经网络的实际推广能力之间往往并不存在严格的定量关系,这就为我们在构造实际的网络中带来了理论与实际的困难。Vapnik V N 和 Chervonenkes A Ja 在他们所提出的 VC 维[1]的基础上,于 1989 年提出了基于结构风险最小化原则的支持向量机[2],有效地将结构风险最小化原则用于统计学习之中。在支持向量机中,通过最小化由结构风险和经验风险组合而成的目标函数,使得学习过程中可能出现欠学习或过学习现象得到了有效的抑制,从而提高了支持向量机的推广能力。经过十几年的发展,支持向量机在模式识别、回归分析和密度估计等方面已经显示出良好的性能。目前,支持向量机已经成为神经网络的一个重要分支,也是从事模式识别和数据挖掘等领域的重要研究手段。

但是,现有的支持向量机,在特征空间中,与所有的支持向量与样本之间的内积相对应的核函数的参数是相同的。这样的支持向量机就只能通过各个支持向量所在的位置及其所对应的权值来描述数据集的分布特征。显然,这样的支持向量机是不能有效地描述不同子区域分布特性显著不同的数据集的分布特征的。这一问题已经引起了人们的注意。对此,人们进行了一些相关的研究[3-5]。然而,目前主要的解决方法是基于划分 - 组合的指导思想。但是由于这类方法存在以下问题:

1)没有整体的风险评估函数,即不能有效地防止欠学习或过学习。

2)邻域的划分,对经验的依赖性较大。

3)不能很好地解决各个邻域之间的边界问题。

为了有效地克服上述困难,笔者提出了一种"基于正反馈的支持向量机"。该支持向量机具有整体上的基于结构风险最小化原则的风险评估函数。通过对人工数据和实际数据的实验,表明基于正反馈的支持向量机在推广性能方面明显优于现有的支持向量机。

1 基于高斯核的支持向量机学习原理

在模式识别中,支持向量机是一种基于结构风险最小化原则,以构造最优超平面为目标的统计学习机器。在以下的论述中,假设训练数据集为:

$(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l),\quad \boldsymbol{x}\in R^n, y\in\{+1,-1\}$

对于在输入空间中不能正确分类的数据集就利用非线性变换 $\Phi(\boldsymbol{x}_i), i=1,2,\cdots,l$,将样本映射到某一更高维的特征空间中,使样本在这个高维的特征空间中能

* 收稿日期:2003 - 12 - 15
作者简介:杨强(1972 -),男,重庆人,重庆大学博士研究生,主要研究方向:支持向量机和图像处理。

尽可能地实现正确分类。在特征空间中，样本之间的内积用核函数 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示。目前，使用得最多的是高斯核函数：

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}) \tag{1}$$

在以下的讨论中，如果没有特别的说明，将用 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 表示高斯核函数。为防止由于高斯核函数的引入，导致过大的结构风险，一般是通过引入正则项，来实现结构风险和经验风险之间的折中。目前，使用较多的是 L1 - SVM 支持向量机，其目标函数为[6]：

$$\begin{aligned}&\min \quad \frac{1}{2}\boldsymbol{\omega}\cdot\boldsymbol{\omega}+C\sum_{i=1}^{l}\xi_i\\&\text{s.t.} \quad \xi_1\geqslant 0,\quad i=1,2,\cdots,l,\\&y_i(\boldsymbol{\omega}\cdot\Phi(\boldsymbol{x}_i)-b)\geqslant 1-\xi_i\end{aligned} \tag{2}$$

其中，C 为一个大于 0 的常数。根据 K - T 条件，得到对偶目标函数：

$$\begin{aligned}&\max \quad W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)\\&\text{s.t.}\ 0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,l\\&\sum_{i=1}^{l}y_i\alpha_i=0\end{aligned} \tag{3}$$

通过求解这个二次规划问题得到相应的系数 $\alpha_i^0\geqslant 0, i=1,2,\cdots,l$，其中大于 0 的系数所对应的样本向量就是所谓的支持向量。进而得到关于最优超平面的权值 $\boldsymbol{\omega}_0$ 和阈值 b_0：

$$\boldsymbol{\omega}_0=\sum_{\text{支持向量}}y_i\alpha_i^0\Phi(\boldsymbol{x}_i) \tag{4}$$

$$b_0=\frac{1}{2}[(\boldsymbol{\omega}_0\cdot\Phi(\boldsymbol{x}^*(1)))-(\boldsymbol{\omega}_0\cdot\Phi(\boldsymbol{x}^*(-1)))] \tag{5}$$

其中，$\boldsymbol{x}^*(1)$，$\boldsymbol{x}^*(-1)$ 分别为属于“+1”类和“-1”类的支持向量，且其对应的 ξ 等于 0。因此，得到基于最优超平面的分类规则的指示函数：

$$f(\boldsymbol{x})=\operatorname{sgn}\Big(\sum_{\text{支持向量}}y_i\alpha_i^0k(\boldsymbol{x}_i,\boldsymbol{x})-b_0\Big) \tag{6}$$

以上是对基于高斯核函数的支持向量机的一般性讨论。该支持向量机具有学习过程简单和能从整体上实现风险最小化等优点，因而目前得到了广泛的研究和应用。但是，这类支持向量机中的所有支持向量与样本所组成的高斯核函数的参数均为 $2\sigma^2$，因此这类支持向量机就只能依靠支持向量的分布和支持向量所对应的权值来对数据的分布特征进行描述。这样的支持向量机在描述不同子区域分布特征显著不同的数据集中，就有可能陷入结构风险与经验风险均难以控制的困境，从而不能实现对这类数据集的有效描述，导致所得到的支持向量机推广性能的比较差。为了解决这一问题，笔者提出基于正反馈的支持向量机。

2 基于正反馈的支持向量机与算法

基于正反馈的支持向量机的学习分两步来实现：

第 1 步，利用 L1 - SVM 支持向量机进行学习。

第 2 步，在第 2 步的学习过程中所使用的核函数如下

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\Big(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^3}{\sigma_i^2+\sigma_j^2}\Big) \tag{7}$$

其中，σ_i^2 表示样本 $\boldsymbol{x}_i$ 与其他样本进行核函数计算（特征空间中的内积）时所对应的核函数参数中由样本 $\boldsymbol{x}_i$ 所确定的部分。同样，σ_j^2 表示样本 $\boldsymbol{x}_j$ 与其他样本进行核函数计算时所对应的核函数参数中由 $\boldsymbol{x}_j$ 所确定的部分。利用 L1 - SVM 中的支持向量分布的特征，通过减小相互间的距离小的支持向量所对应的核函数参数 σ^2（并且使学习样本集中的非支持向量的学习样本所对应的核函数参数 σ^2 与离它最近的支持向量所对应的核函数参数 σ^2 相等），使得通过学习后所得到的位于该区域附近的支持向量所对应的幅值递减得更快，以提高基于正反馈的支持向量机在该区域的灵敏度；通过增大相互间的距离大的支持向量所对应的核函数参数 σ^2（并且使学习样本集中的非支持向量的学习样本所对应的核函数参数 σ^2 与离它最近的支持向量所对应的核函数参数 σ^2 相等），使得通过学习后所得到的位于该区域附近的支持向量所对应的幅值递减得更慢，以减少基于正反馈的支持向量机在该区域的结构风险。这也正是将该类支持向量机命名为基于正反馈的支持向量机的原因所在。

在上述分析的基础上，提出基于正反馈的支持向量机的算法：

1）用 L1 - SVM 支持向量机模型对学习样本进行学习，对于大样本集，可以采用序列化算法进行学习[7]。得到 L1 - SVM 支持向量机。

2）在输入空间中，计算 L1 - SVM 支持向量机中每个支持向量 $\boldsymbol{x}_i$ 与离该支持向量最近的支持向量之间的距离 d_i。

3）计算 L1 - SVM 支持向量机中所有支持向量所对应的距离 d_i 的平均值 d。

4）用式（8）得到 L1 - SVM 支持向量机中每个支持向量所对应的核函数参数 σ_i^2

$$\sigma_i^2=\frac{d_i}{d}\times\sigma_2 \tag{8}$$

5）对于用 LI - SVM 学习后，学习样本集中的各个

非支持向量的学习样本，找出与之最近的 L1 - SVM 支持向量机中的支持向量 x_i，则该学习样本所对应的核函数参数为 σ_i^2。

6)对所有的学习样本(无论是否为 L1 - SVM 支持向量机中的支持向量)计算核函数

$$k(x_i,x_j) = \exp\left(-\frac{\|x_i - x_j\|^3}{\sigma_i^2 + \sigma_j^2}\right)$$

7)在核函数为式(7)的情况下，求解最优目标函数式(3)，得到基于正反馈的支持向量机中的支持向量和相应的系数 α_i^0。

8)利用式(9)计算基于正反馈的支持向量机的阈值

$$b_0 = \frac{1}{n_1}\sum_{j1}\left(\sum_i \alpha_i^0 y_i \exp\left(-\frac{\|x_i - x_{j1}\|^3}{\sigma_i^2 + \sigma_{j1}^2}\right)\right) - \frac{1}{n_2}\sum_{j2}\left(\sum_i \alpha_i^0 y_i \exp\left(-\frac{\|x_i - x_{j2}\|^3}{\sigma_i^2 + \sigma_{j2}^2}\right)\right) \quad (9)$$

其中 i 表示基于正反馈的支持向量机中的所有支持向量，n_1 表示属于“+1”类的支持向量的个数，$j1$ 表示属于“+1”类的支持向量，且对应的等于 0 支持向量。n_2 表示属于“-1”类的支持向量的个数，$j2$ 表示属于“-1”类的支持向量，且对应的 ξ 等于 0 的支持向量。

9)基于正反馈的支持向量机的判别函数为

$$f(x) = \mathrm{sgn}\left(\sum_{\text{支持向量}} y_i \alpha_i^0 k(x_i,x) - b_0\right) \quad (10)$$

其中，式(10)中的 $k(x_i,x)$ 为

$$k(x_i,x) = \exp\left(-\frac{\|x_i - x\|^3}{2\sigma_i^2}\right) \quad (11)$$

3　实验结果与分析

为了检验基于正反馈的支持向量机的正确性，笔者用人工数据集和现实数据集进行实验。其中人工数据集为双螺旋数据[8]。真实数据用 breast - cancer 数据库[9]。

图 1 是一般的双螺旋图形。因为希望利用双螺旋数据集来检验基于正反馈的支持向量机对不同子区域分布不同的数据集的分布特征的描述能力，所以将一般的双螺旋数据集进行了一定的修改，将数据的发散速度与半径的平方成正比。所得到的数据分布如图 2 所示。

在对如图 1 所示的双螺旋数据集进行学习之前，将数据集中的所有数据间次地分为学习样本集和测试样本集。这样就满足了学习样本集与测试样本集的独立同分布条件。L1 - SVM 与基于正反馈的支持向量机对如图 2 所示的双螺旋数据集的实验结果如表 1 所示。

其中，P - SVM 表示基于正反馈的支持向量机，numSVM 表示支持向量的个数，Bias 表示支持向量机

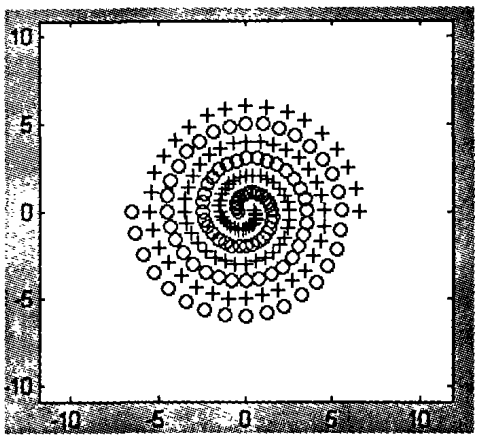

图 1　双螺旋

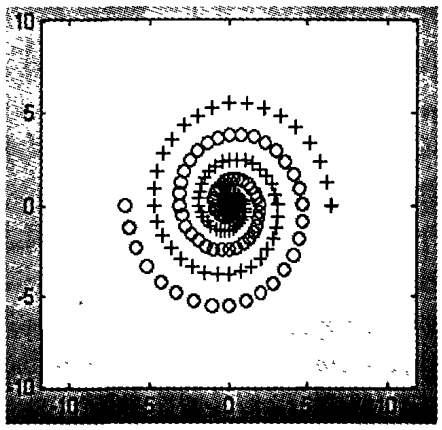

图 2　双螺旋改进图

的阈值，E% 表示对测试样本的错误率。

表 1　对双螺旋数据集的实验结果

方法	numSVM	Bias	E%
L1 - SVM	85	-0.015 7	6.2
P - SVM	78	0.00 77	4.1

L1 - SVM 与基于正反馈的支持向量机对 breast - cancer 数据库进行实验的结果如表 2 所示。

表 2　对 breast - cancer 数据库的实验结果

方法	numSVM	bias	E%
L1 - SVM	169	-0.023 9	29.87
P - SVM	165	-0.022 4	23.30

从对如图 2 所示的双螺旋数据集和 breast - cancer 数据库进行的对比实验的结果可知，基于正反馈的支持向量机相对于 L1 - SVM 支持向量机而言，能够更好地描述数据集的分布特征，从而得到更好的推广能力。特别值得注意的是，从对 breast - cancer 数据库的对比实验中所得到的结果可知，基于正反馈的支持向量机不仅比 L1 - SVM 支持向量的个数更少，而且所得到的结果明显好于目前其他任何方法，包括其它类型的支持向量机所得到的结果。基于正反馈的支持向量机与其他方法对 breast - cancer 数据库的实验如表 3 所示。其他方法对 breast - cancer 数据库的实验结果

可以从文献[9]得到。

表3 多种方法的实验效果对比

方法	P-SVMR	BF-Network	AdaBoost	SVM	Kernel Fisher
E%	23.30	27.64	30.36	26.04	24.77

另外,通过两种支持向量机对如图2所示的双螺旋数据集和 breast-cancer 数据库进行的实验得知,P-SVM中的支持向量的个数均比 L1-SVM 中的支持向量的个数少。这说明基于正反馈的支持向量机确实是一种比较好的支持向量机。

4 结 论

在分析当前的支持向量机在描述不同子区域分布特征不同的数据集的分布特征时所遇到的困难的基础上,突破在构造支持向量机中"对各个学习样本一定要选择相同的核函数参数"的这一苛刻要求,提出了基于正反馈的支持向量机这一新型支持向量机。应该说这是支持向量机发展中的一个突破。实验证明,基于正反馈的支持向量机能够更好地描述数据集的分布特征。另外,值得说明的是,在本文中虽然只讲述了用于模式识别的基于的正反馈的支持向量机,但是,在基于正反馈的支持向量机中所蕴涵的思想:利用正反馈来建立由核函数参数不同的支持向量地组成支持向量机,对于其他类型的支持向量机也是适用的。

参考文献:

[1] VLADIMIRNVAPNIK, CHERVONENKIS A JA. On the uniform convergence of relative frequencies of events to their probabilities[M]. USSR: Doklady Akademii Nauk, 1968.181.

[2] VLADIMIRNVAPNIK, CHERVONENKIS A JA. The necessary and sufficient conditions for consistency of the method of empirical risk minimization [J]. Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting, 1989, (2): 207-249.

[3] MILIDIU R L, MACHADO R J, RENTERA R P. Time-series forecasting through wavelets transformation and a mixture of expert models [J]. Neurocomputing, 1999, 28: 145-146.

[4] KWOK T J. Support vector mixture for classification and regression problem[A]. ICPR98: Proceedings of the 14th International Conference on Pattern Recognition [C]. Brisbane, Australia: 1998.255-258.

[5] LIJUAN CAO. Support vector machines experts for time series forecasting[J]. Neurocomputing, 2003, 51: 321-339.

[6] VLADIMIR NVAPNIK. Statistical Learning Theory [M]. New York: Wiley, 1999.

[7] JOHN C PLATT. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines in Advances in Kernel Methods-support Vector Learning[M]. Cambridge, MA: MIT Press, 1999, 185-208.

[8] MATT WHITE. CMU Neural Networks Benchmark Collection [EB/OL]. http://www-2.cs.cmu.edu./afs/cs/project/ai-repository/aiareas/neural/bench/cmu, 2003-12-10.

[9] BLAKE, CLMERZ. UCI Machine Learning Repository[EB/OL]. http://ida.first.gmd.de/~raetsch/data, 2003-07-08.

Support Vector Machines Based on Positive Feedback

YANG Qiang, WU Zhong-fu, YU Ping, ZHONG Jiang

(College of Computer, Chongqing University, Chongqing 400030)

Abstract: Support vector machines based on positive feedback are put forward with the analysis of both advantages and disadvantages of current support vector machines based on Gaussian kernel function (including support vector machines based on K-nearest neighbors) and the essence why support vector machines is capable of describing data sets' distribution characteristics, thus the rigor constraint is overcome that maintains "corresponding parameters of kernel function support vectors should be equal". The learning algorithm of support vector machines based on positive feedback is given. Simulation experiments of artificial and real data proves that support vector machines based on positive feedback is be obviously superior to current ones in its generation capabilities. Though, only the support vector machines based on positive feedback for pattern recognition is discussed, the idea included in support vector machines based on positive feedback is using kernel functions with different corresponding parameters to construct support vector machines, and is adaptive to other types support vector machines.

Key words: RBF; K-nearest neighbors; kernel function; support vector machines

(编辑 张 苹)