支持向量回归中的预测信任度*)

孙德山[1]　吴今培[2]

（中南大学科研所　长沙410075）[1]　（五邑大学智能所　广东江门529020）[2]

Predicting Credibility Based on Support Vector Regression

SUN De-Shan[1]　WU Jin-Pei[2]

(Department of Scientific Research, Central South University, Hunan,Changsha, 410075)[1]

(Institute of Intelligence Technology & System Wuyi University, Guangdong, Jiangmen, 529020)[2]

Abstract Support vector machine(SVM)has been widely applied to classification and regression problems, but it suffers from some important limitations, one of the most significant being that it makes point predictions rather than generating probability output. A notion of predicting credibility is proposed in support vector regression machine based on the problem, which can make predicting value have a definite measure, and then relationship between predicting credibility and noise is discussed. Finally, an example of predicting chaotic time series shows the rationality of the definition.

Keywords Support vector machine ,Regression ,Chaotic time series

1 引言

Vapnik 等人根据统计学习理论提出的支持向量机学习方法[1]，近年来受到了国际学术界的广泛重视。支持向量机的最大特点是根据 Vapnik 结构风险最小化原则，尽量提高学习机的泛化能力，即由有限的训练集样本得到的小的误差能够保证对独立的测试集仍然保持小的误差。另外，由于支持向量机算法是一个凸优化问题，因此局部最优解一定是全局最优解。这些特点是其它学习算法，如神经网络学习算法所不及的。但在支持向量回归预测中，它也有一定的不足，比如，它的预测输出是一点，没有概率特征，这就使得它的预测没有一定的依靠性，使得人们不知道这个预测是否可信。针对这点，本文在支持向量机的回归算法中引入了预测信任度的概念，对预测输出的值，能够清楚地看到它的可信度大小，从而能够指导我们对预测值的取舍。文中最后给出具体实例来说明它的合理性。

2 支持向量机的回归算法[2~5]

设给定的训练样本为$\{(x_1,y_1),\cdots,(x_l,y_l)\}\subset R^n\times R$，首先使用一非线性映射 ϕ 把数据映射到一个高维特征空间，再在高维特征空间进行线性回归。设线性函数为：

$$f(x)=\langle w,\phi(x)\rangle+b$$

优化问题是最小化

$$\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$

条件为

$$f(x_i)-y_i\leqslant\xi_i^*+\varepsilon,i=1,\cdots,l$$

$$y_i-f(x_i)\leqslant\xi_i+\varepsilon,i=1,\cdots,l$$

$$\xi_i,\xi_i^*\geqslant0,i=1,\cdots,l$$

优化式子中第一项使函数更为平坦，从而提高泛化能力，第二项则为减小误差，常数 C 对两者做出折中。ε 为一正常数，$f(x_i)$与 y_i 的差别小于 ε 时不计入误差，大于 ε 时误差计为$|f(x_i)-y_i|-\varepsilon$。

这也是一个凸二次优化问题，通过求其对偶问题，得到最大化下面函数：

$$W(\alpha,\alpha^*)=-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle\phi(x_i),\phi(x_j)\rangle+\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i-\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon$$

其约束为

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0$$

$$0\leqslant\alpha_i,\alpha_i^*\leqslant C,i=1,\cdots,l$$

这也是一个二次优化问题，w 由下式给出

$$w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\phi(x_i)$$

函数 $f(x)$可表示为：

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\langle\phi(x_i),\phi(x)\rangle+b$$

由于在上面的式子中只考虑到高维特征空间中的点积运算，因此只要找到一个核函数 $K(x,y)=\langle\phi(x),\phi(y)\rangle$就可以了。已经证明，对称函数 $K(x,y)$只要满足 Mercer 条件即可满足要求，这样就避免了明确知道 $\phi(x)$，巧妙地解决了在高维空间中的运算。

按照 Kuhn-Tucker 定理，可得 b 的计算式如下

$$b=y_j-\varepsilon-\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)K(x_j,x_i)\quad 当\ \alpha_i\in(0,C)$$

$$b=y_j+\varepsilon-\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)K(x_j,x_i)\quad 当\ \alpha_i^*\in(0,C)$$

3 预测信任度

用上面支持向量回归算法建立起预测模型，通常把待预测样本输入，只能获得一个点输出，至于这个预测值到底可信与否并不知道。本文根据局部预测的思想，对每个待预测样本都找到它的一个邻近集，然后以这个邻近集作为训练集，并且

*）广东省自然科学基金资助(021349)。孙德山　讲师，博士研究生，主要研究方向：支持向量回归、时间序列分析。吴今培　教授，博士生导师，主要研究方向：人工智能及电子技术。

根据这个邻近集来定义该待预测样本的预测信任度。这样不仅使训练样本质量高而且数量少，从而提高程序运行速度，而且对每个待预测样本都能提供一个预测信任度。

设总的训练集为$\{x_i, y_i\}_{i=1}^{M}$，M 表示训练数据的个数，$x_i \in R^m$ 是输入数据，$y_i \in R$ 是输出数据。目的就是给出待预测样本 x_{M+1}，获得预测值 $\hat{y}_{M+1}$。根据前面的思想，首先选择待预测样本的邻近集，给定一个适当的正数 δ，选择到 x_{M+1}的距离小于等于 δ 的样本作为训练集（设数量为 l），然后送入支持向量回归机中学习，得到 $f(x)$。接下来，给出能够接受的误差 $e>0$，预测值 $\hat{y}_{M+1}$的 e-预测信任度定义为

$$PB_e(\hat{y}_{M+1}) = \frac{\sum_{i=1}^{l} I\{|y_i - f(x_i)| \leq e\}}{l},$$

其中，$I(A)=\begin{cases}1, 若\ A\ 为真 \\ 0, 否则\end{cases}$

上面的定义也可以被理解为预测值$\hat{y}_{M+1} \in [y_{M+1}-e, y_{M+1}+e]$的概率为上面定义的预测信任度，这就使我们对预测值的可信度有了一定的了解。可信度随着 e 的变化而变化，如果当 e 相对比较小时，可信度还很高的话，说明数据中几乎没有噪声，相反，可信度比较小的话，说明数据中可能含有一定的噪声。从直观上看，e 和可信度之间的关系能够反映数据中所含噪声的多少。当然，可信度的大小还和支持向量回归机损失函数中的 ε 值有关，ε 值越小，可信度应该越大。而 e 值越小，可信度就越小，由此可见 ε 和 e 对可信度的影响成相反关系。若取 $e=\varepsilon$，当 ε 值逐渐减小时，可信度的变化相对比较平稳，则说明数据中噪声很小或没有噪声，若可信度变化较快，则说明数据中含有噪声。总之，通过对各参变量的调整，能够在一定程度上辨别出数据中噪声的含量。

4 举例

下面以一个混沌时间序列预测为例来说明本文所定义的预测信任度是合理性的。Hénon 混沌时间序列是由下面的式子产生的：

$$X_{n+1}=1-1.4Xn^2+Y_n, Y_{n+1}=0.3X_n$$

取初值[0, 0]，取其中的一维得到288个样本，前280个作为训练样本，后8个作为测试样本，为了对该时间序列进行预测，首先要重构相空间，取重构维数 $m=2$，其他参数取为 $\delta=0.2$，$\varepsilon=0.001$，$e=0.001$，$C=500$。首先是不含噪声的数据，所得结果如表1；加入 $N(0,0.001)$的噪声，所得结果如表2；加入 $N(0,0.01)$的噪声，所得结果如表3。

表1 不含噪声

实际值	1.2110	−1.0020	−0.0423	0.6969	0.3074	1.0768	−0.5311	0.9282
预测值	1.2110	−1.0019	−0.0422	0.6969	0.3075	1.0768	−0.5312	0.9282
信任度	1	1	1	1	1	1	1	1

表2 加入 $N(0,0.001)$的噪声

实际值	1.2106	−1.0019	−0.0427	0.6963	0.3083	1.0760	−0.5313	0.9287
预测值	1.2112	−1.0022	−0.0413	0.6965	0.3058	1.0756	−0.5287	0.9284
信任度	0.6111	0.45	0.6316	1	0.7143	0.5152	0.32	0.6

表3 加入 $N(0,0.01)$的噪声

实际值	1.2216	-1.0044	-0.0575	0.6970	0.3081	1.0800	-0.5261	0.9410
预测值	1.2234	-0.9994	-0.0536	0.6813	0.2824	1.0699	-0.5520	0.9545
信任度	0.3889	0.05	0	0	0.6667	0.2813	0.0385	0.3

从上面例子中，我们能清楚地看到，每预测一个值时，都能给出一个信任度，这样就给预测提供了一定的依据，避免了盲目性。另外，从信任度的大小可以看出数据中的噪声含量，如表3中的预测值与实际值也比较接近，但预测信任度却很小，说明数据中含有一定的噪声，并且可以初步断定噪声的方差应该大于给定的接受误差 e 的值。还可以适当调整参数和的大小，看信任度的变化情况，来进一步分析数据中的噪声含量。这个例子说明了文中所定义的预测信任度的合理性以及与噪声之间的关系。

结论 支持向量机是一种基于统计学习理论的新颖的机器学习方法，由于其出色的学习性能，该技术已成为当前国际机器学习界的研究热点。支持向量机算法找到的是全局最优解，因此，在很多问题上它都有着其他统计学习技术所难以比拟的优越性，并在很多领域都得到了成功的应用，如人脸检测、手写体数字识别、文本自动分类等。但它还有一些不足之处，比如说，其中的参数选择问题，核函数选择问题，还有就是支持向量机只有点输出，没有概率输出。文[6,7]中提出一些办法，比如引入贝叶斯方法到支持向量机中，赋予它一定的概率特性，但这些方法的给出也同时提出了新的问题，比如先验概率比较难确定，程序复杂等。本文在支持向量回归中定义了一种预测信任度的概念，使得预测值有了一个可信的量度，并且能够在一定程度上分析数据中噪声的含量。由于对每个待预测样本都采用局部邻近训练集，这样不仅提高了预测精度，而且还能够提高程序运行速度。

参考文献

1 Vapnik V N. Statistical Learning Theory [M]. New York, Wiley, 1998

2 Smola A J, Scholkopf B. A tutorial on support vector regression [R]: [NeuroCOLT TR NC-TR-98-030]. Royal Holloway College University of London, UK, 1998

3 Fernandez R. Predicting time series with a local support vector regression machine [C]. IN ACAI 99, 1999

4 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition [R]. Knowledge Discovery and Data Mining, 1998, 2 (2)

5 Cortes C, Vapnic V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273～297

6 MacKay D J C. Bayesian interpolation [J]. Neural Computation, 1992, 4(3): 415～447

7 MacKay D J C. The evidence framework applied to classification networks [J]. Neural Computation, 1992, 4: 720～736