基于支持向量回归的唇动参数预测

王志明 蔡莲红 艾海舟
(清华大学计算机科学与技术系 北京 100084)
(wzm00@mails.tsinghua.edu.cn)

摘 要 支持向量机学习方法以结构风险最小化原则取代传统机器学习方法中的经验风险最小化原则,在有限样本的机器学习中显示出优异的性能.将这一新的统计学习方法应用到多媒体交互作用的研究中,用支持向量回归的方法由语音预测唇动参数.通过对语音的线性预测系数进行主分量分析,有效地压缩了声学特征参数的维数.结合交叉校验和最速下降优化方法,选择最佳的支持向量回归学习参数.在汉语0~9的任意数字串上对唇高参数的预测实验结果达到了均方误差0.0096,平均幅度误差7.2%及相关系数0.8的效果.这一结果优于一个文中优化过的人工神经网络所达到的性能,说明这一方法很有潜力.

关键词 支持向量机;支持向量回归;线性预测系数;主分量分析;人工神经网络

中图法分类号 TP391

Mouth Movement Prediction Based on Support Vector Regression

WANG Zhi-Ming, CAI Lian-Hong, and AI Hai-Zhou
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Unlike traditional machine learning which is based on empirical risk minimization principle, support vector machine (SVM) learning is based on structural risk minimization principle. SVM shows powerful ability in learning with limited samples. This new method is applied in the study of multimedia interaction and in predicting the mouth movement by speech based on support vector regression (SVR). The audio parameters dimension is reduced by principle components analysis (PCA), and the optimal SVR learning parameters are selected based on cross-validation and steepest descent algorithm optimization. With the experiment on arbitrary Chinese digital numbers from 0 to 9, the prediction results reach 0.0096 in mean square error, 7.2% in absolute magnitude error, and 0.8 in linear correlation coefficient. It gives better results than that with optimized artificial neural network, which shows that the proposed method is quite promising.

Key words support vector machine (SVM); support vector regression (SVR); linear predictive coding (LPC); principal components analysis (PCA); artificial neural network (ANN)

1 引 言

随着多媒体技术的飞速发展,多种媒体之间的交互作用正越来越受到人们的普遍关注.语音和图像是人们日常交流中最常用的传输媒体,挖掘二者之间的内在联系如语音与唇动之间的关系有着广泛的应用前景.如在网络传输多媒体数据时往往会受到网络带宽的限制,无法同时提供语音信号与视频信号.如果我们可以从音频数据中预测出相应的唇

收稿日期:2002-09-19;**修回日期**:2003-03-31
基金项目:高等学校博士学科点专项科研基金(20010003049)

动参数,则可以利用计算机动画的方法合成人脸动画,改善人机交互界面并增强人们对语音的理解;利用语音与唇动之间的相关性,还可实现视频会议等场合下的视频流丢帧插补、多模态的身份验证等.

如果依靠语音识别来识别出相应的文本,再合成人脸动画,存在较大的问题. 首先,语音识别需要较长的时延,识别系统往往要等到人们说完一句话才能得出识别结果;其次,现在语音识别系统对任意文本、连续语流的识别率较差;第 3,从语音到文本的过程会丢失语音中许多有用的信息,如音量的大小,说话者的情感色彩等等. 而直接从语音预测唇动参数则可以有效地保留这些信息,且时延很小(与预测时考虑的前向语音帧数有关).

在由语音预测唇动参数的研究中,AT&T Bell 实验室的 Tsuhan Chen 和 Rao 等人做了长期的、大量的研究工作,尝试了各种预测方法,包括基于矢量量化(vector quantization)分类的方法、基于人工神经网络(artificial neural network)的方法、基于混合高斯模型(Gauss mixed model)的方法、基于隐马尔克夫模型(hide Markov model)的方法等[1~4]. Lavagetto 采用了时延神经网络(time-delay neural network)的方法[5], Choi 和 Williams 等人也采用了基于隐马尔克夫模型的方法[6,7]. 但这些方法都具有一定的局限性:基于分类的方法会产生不连续的预测结果;神经网络易产生过学习现象;基于混合高斯模型的统计方法在训练样本不足时会产生较大的误差;而基于隐马尔克夫模型的方法一般需要较长的时延. 本文尝试用一种新的统计学习方法——支持向量回归(support vector regression)——来进行唇动参数的预测.

支持向量机(support vector machine, SVM)是统计学习理论在近些年发展出的一种新的机器学习方法. 它建立在一套坚实的理论基础之上,它的指导原则是同时优化经验风险和模型复杂度,在解决有限样本学习问题时表现出优异的性能. 我们在应用支持向量回归进行唇动参数预测时,通过交叉校验(cross validation)和梯度下降的方法获取最佳的支持向量回归的学习参数. 实验表明,它的性能优于本文经过仔细优化的神经网络.

本文的内容安排如下:第 2 节介绍了支持向量回归的基本原理和方法;第 3 节阐述了基于支持向量回归的唇动参数预测方法;第 4 节是实验结果,本文同时给出了神经网络的实验结果以便比较;最后是结束语.

2 支持向量回归的优势

给出一组实验数据$(y_i, \boldsymbol{x}_i)$, $i=1,2,\cdots,l$,为了得出输入 $\boldsymbol{x}_i$ 和输出 y_i 之间的关系,传统的回归学习方法(如神经网络)在经验风险最小化原则 ERM (empirical risk minimization),即最小化

$$R_{\text{emp}}[f] = \frac{1}{l}\sum_{i=1}^{l} L(f(\boldsymbol{x}_i) - y_i) \tag{1}$$

的条件下求最佳的回归函数 $y = f(\boldsymbol{x})$,其中 $L(f(\boldsymbol{x}_i) - y_i)$ 表示实际值为 y_i,回归函数值为 $f(\boldsymbol{x}_i)$所造成的误差函数值.

而统计学习理论指出,在有限样本的情况下,经验风险最小并不能保证实际风险最小,典型的情况是神经网络的过学习现象. 支持向量回归是基于结构风险最小化原理 SRM(structural risk minimization principle)的学习方法,它同时最小化经验风险和模型复杂度,保证了有限样本情况下模型的最佳推广能力和输出函数的平滑性.

支持向量回归可表示如下[8]:

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(\boldsymbol{x}_i, \boldsymbol{x}) + b = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}) + b = \boldsymbol{w}\phi(\boldsymbol{x}) + b, \tag{2}$$

其中,$\alpha_i^* \geqslant 0$, $\alpha_i \geqslant 0 (i=1,2,\cdots,l)$,不等于零的项所对应的 $\boldsymbol{x}_i$ 即为支持向量,$K(\boldsymbol{x}_i, \boldsymbol{x}) \equiv \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x})$ 表示所采用的点积核函数, $\boldsymbol{w} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\phi(\boldsymbol{x}_i)$.

求解系数 α_i^*, α_i 和 b 的过程为最小化式(3):

$$\frac{1}{2}\| \boldsymbol{w} \|^2 + C \cdot R_{\text{emp}}[f]. \tag{3}$$

式(3)第 1 项 $\| \boldsymbol{w} \|^2$ 定义了模型复杂度,第 2 项表示经验损失函数,C 为一平衡常数. 在损失函数采用 ε 不敏感函数

$$L(f(\boldsymbol{x}_i) - y_i) = | y - f(\boldsymbol{x}) |_\varepsilon = \max\{0, | y - f(\boldsymbol{x}) | - \varepsilon\},$$

并引入一对松弛项 ξ_i 和 ξ_i^* 后,式(3)可表示在约束条件

$$(\boldsymbol{w} \cdot \boldsymbol{x}_i) + b - y_i \leqslant \varepsilon + \xi_i,$$
$$y_i - (\boldsymbol{w} \cdot \boldsymbol{x}_i) - b \leqslant \varepsilon + \xi_i^*,$$
$$\xi_i > 0, \ \xi_i^* > 0, \ \varepsilon > 0$$

下极小化

$$\tau(\boldsymbol{w}, \xi, \xi^*) = \frac{1}{2}\| \boldsymbol{w} \|^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*). \tag{4}$$

引入 Lagrange 乘子 α 和 α^* 后,根据 Wolfe 对偶原理,可等价于在约束条件 $\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0$, $\alpha_i \in [0, \frac{C}{l}]$, $\alpha_i^* \in [0, \frac{C}{l}]$ 下极大化式(5):

$$W(\alpha, \alpha^*) = -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\boldsymbol{x}_i \cdot \boldsymbol{x}_j). \quad (5)$$

通过极大化式(5)得到最优的 α_i^*, $\alpha_i (i = 1, 2, \cdots, l)$ 后,利用任意训练数据,由 $y_i = \boldsymbol{w}\phi(\boldsymbol{x}_i) + b$ 可求得参数 b,这样便可由式(2)对新的输入数据进行预测.

Bernhard Schölkopf 等人于 1998 年提出了改进的 μ-SVR 方法[9],它将式(4)的极小化变为

$$\tau(\boldsymbol{w}, \xi, \xi^*, \varepsilon) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\mu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right). \quad (6)$$

它通过增加一个新的常数 μ 和作为变量的 ε 调节模型复杂度和松弛变量. 同样根据对偶原理,可以转化为在约束条件 $\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0$, $\alpha_i \in [0, \frac{C}{l}]$, $\alpha_i^* \in [0, \frac{C}{l}]$, $\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) \leqslant C \cdot \mu$ 下极大化式(7):

$$W(\alpha, \alpha^*) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\boldsymbol{x}_i \cdot \boldsymbol{x}_j). \quad (7)$$

Bernhard Schölkopf 指出这一算法可自动最小化 ε,还证明了常量 μ 是错误率的上界和支持向量比率的下界. 本文在由语音预测唇动参数的过程中采用 μ-SVR 方法,并根据交叉校验的方法确定了最优的常数项 C, μ 等参数.

3　利用支持向量回归进行唇动参数预测

为了从语音信号预测说话者的唇部运动参数,本文首先选择了有效的声学特征参数,然后在用支持向量回归进行唇动参数的预测中,优化支持向量回归的几个学习参数,如常数 C, μ 等.

3.1　语音参数的选择

通过实验对比语音的线性预测系数 LPC(linear predictive coding)、线谱频率系数 LSF(line spectral frequency)、实倒谱系数 RCEP(real cepstrum)、反射系数 RC(reflection coefficient)和 Mel 倒谱系数 MFCC(Mel frequency cepstrum coefficient)参数发现,在进行唇动参数预测中,最为有效的是 LPC 参数. 语音短时能量和过零率对预测的准确性也有一定的帮助.

由于协同发音的影响,每一时刻的口形不只与当前所发的语音有关,还与其前后一定时间范围内的语音有关. 实验发现这一范围约为 150～200 ms,且当前时刻之后的语音比当前时刻之前的语音影响更大. 如果以 30 ms 每帧,20 ms 帧移计算语音参数,取前 2 帧、后 4 帧和当前帧的共计 7 帧语音参数来预测当前唇动参数,每帧语音计算其 8 阶 LPC 参数、语音短时能量和过零率,这样总的参数将高达 70 维. 在学习过程中增加了模型的复杂度,需要较大的计算量,为了有效地压缩输入参数的维数,本文对语音参数进行主分量分析(PCA),即进行了 K-L 变换. 分析显示各维参数之间存在较强的相关性,变换后的各维数据所占的信息比和累积信息量如表 1 所示:

表 1　语音参数经 K-L 变换后各维所占信息比和累积量

维度	信息比	累积量	维度	信息比	累积量
1	90.23	90.23	6	0.30	99.62
2	7.00	97.23	7	0.21	99.83
3	0.89	98.12	8	0.12	99.95
4	0.78	98.90	9	0.03	99.98
5	0.42	99.32	10	0.02	100.00

从表 1 可以看出,取前 4 个主分量已可以包含 98.9% 的有用信息. 因此,本文选择语音参数经 K-L 变换后的前 4 个主分量进行唇动参数预测,这样输入参数的维数降为 $4 \times 7 = 28$ 维.

3.2　支持向量回归学习参数的选择

在应用 μ-SVR 进行唇动参数的预测中,本文选用了高斯径向基函数(Gaussian radial basis function)作为点积核函数,即

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}|^2}{2\sigma^2}\right). \quad (8)$$

训练需要预先确定学习参数 C, μ, σ,本文采用交叉校验(cross validation)的方法来确定这两个参数. 将训练集分为 10 等份,每次留一份验证模型性能,其余 90% 用来训练模型,最后以模型在 10 次验证数据上的性能平均值作为这一学习参数下的模型性能.

将参数 C, μ, σ 作为变量,用最速下降优化方

法来极小化 SVR 模型性能(MSE)的具体算法如下:

(1) 设置初始的参数值(C,μ,σ)和参数改变的步进量 $\Delta C,\Delta\mu,\Delta\sigma$,置迭代次数 $Iters=1$,性能没有改善的连续迭代次数 $Nochange=0$;

(2) 计算当前参数值下交叉校验的平均性能 $Mse(C,\mu,\sigma)$.

(3) 将参数值分别更改为($C-\Delta C,\mu,\sigma$)和($C+\Delta C,\mu,\sigma$),计算相应的平均性能 $Mse(C-\Delta C,\mu,\sigma)$和 $Mse(C+\Delta C,\mu,\sigma)$,若 $Mse(C-\Delta C,\mu,\sigma)$和 $Mse(C+\Delta C,\mu,\sigma)$均大于 $Mse(C,\mu,\sigma)$,则记 $\Delta Mse_C=0$;否则,若 $Mse(C-\Delta C,\mu,\sigma)<Mse(C+\Delta C,\mu,\sigma)$,记 $\Delta Mse_C=Mse(C-\Delta C,\mu,\sigma)-Mse(C,\mu,\sigma)$,反之,记 $\Delta Mse_C=Mse(C,\mu,\sigma)-Mse(C+\Delta C,\mu,\sigma)$;

(4) 与(3)类似将参数值分别更改为($C,\mu-\Delta\mu,\sigma$)和($C,\mu+\Delta\mu,\sigma$),并计算出 ΔMse_μ;将参数值分别更改为($C,\mu,\sigma-\Delta\sigma$)和($C,\mu,\sigma+\Delta\sigma$),并计算出 ΔMse_σ;

(5) 令 $\Delta Mse=\max(\mathrm{abs}(\Delta Mse_C,\Delta Mse_\mu,\Delta Mse_\sigma))$,若 $\Delta Mse=0$,令 $Nochange=Nochange+1$,同时令 $\Delta C=\Delta C/2,\Delta\mu=\Delta\mu/2,\Delta\sigma=\Delta\sigma/2$;否则令 $C=C+\Delta C\cdot\Delta Mse_C/\Delta Mse$, $\mu=\mu+\Delta\mu\cdot\Delta Mse_\mu/\Delta Mse$, $\sigma=\sigma+\Delta\sigma\cdot\Delta Mse_\sigma/\Delta Mse$, $Nochange=0$;

(6) $Iters=Iters+1$,若 $Iters$ 大于规定的最大迭代次数或 $Nochange$ 大于允许的最大无性能改善迭代次数,则退出;否则返回(2).

经上述迭代过程得到最优的 C,μ,σ 后,以这些参数在所有的训练数据上训练得到一个性能最优的 μ-SVR 模型.

4 实验结果

本文的实验语料为 100 遍重复的 0~9 数字串,实验数据为单个女性发音人对上述语料的发音,总的发音长度约473 s,语音采样频率为11.025 kHz,语音帧长为30 ms,20 ms帧移. 图像分辨率 640×480,图像帧速率为25 fps,共计 11814 个样本.

为评价回归学习的质量,本文计算了均方误差 MSE(mean square error)、平均绝对幅度误差 MAE(mean absolute error)和线性相关系数 R(linear correlation coefficients). 设唇动参数的真实值均值为 $\mu_y=\frac{1}{l}\sum_{i=1}^{l}y_i$,参数变换幅度范围为 $\Delta=\max(y)-\min(y)$,预测值的平均值为 $\mu_f=\frac{1}{l}\sum_{i=1}^{l}f(\boldsymbol{x}_i)$,则各参数的计算公式如下:

均方误差 MSE:

$$mse=\frac{1}{l}\sum_{i=1}^{l}\left(\frac{y_i-f(\boldsymbol{x}_i)}{\Delta y}\right)^2. \tag{9}$$

平均绝对幅度误差 MAE:

$$mae=\frac{1}{l}\sum_{i=1}^{l}\left|\frac{y_i-f(\boldsymbol{x}_i)}{\Delta y}\right|. \tag{10}$$

线性相关系数 R:

$$R=\frac{\mathrm{cov}(y_i,f(\boldsymbol{x}_i))}{\sqrt{\mathrm{cov}(y_i,y_i)\cdot\mathrm{cov}(f(\boldsymbol{x}_i),f(\boldsymbol{x}_i))}}=\frac{\sum_{i=1}^{l}(y_i-\mu_y)(f(\boldsymbol{x}_i)-\mu_f)}{\sqrt{\sum_{i=1}^{l}(y_i-\mu_y)^2\cdot\sum_{i=1}^{l}(f(\boldsymbol{x}_i)-\mu_f)^2}}. \tag{11}$$

实验中从所有实验数据中等间隔地取 1/8 作为测试样本,其余 7/8 作为训练数据. 对于唇高参数的实验结果如表 2 所示:

表 2 支持向量回归优化前后及与神经网络的性能比较

预测方法		SVR(参数优化前)	SVR(参数优化后)	ANN
训练集	MSE	0.0127	0.0095	0.0104
	MAE	0.0859	0.0711	0.0758
	R	0.724	0.805	0.784
测试集	MSE	0.0132	0.0096	0.0104
	MAE	0.0873	0.0720	0.0760
	R	0.711	0.0800	0.781

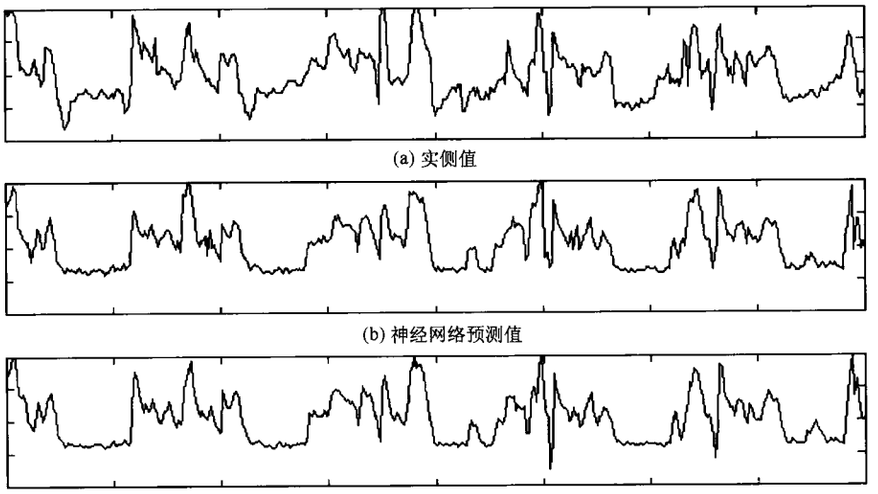
从表 2 可以看出,经交叉校验和梯度下降法优化 SVR 的学习参数后,系统 MSE 性能改善了约 27%(测试集),其中优化前参数为:$C=1$, $\mu=0.5$, $\sigma=0.04$,优化后参数为:$C=4.35$, $\mu=0.54$, $\sigma=0.30$. 另外,从表中还可以看出,经过优化的 SVR 性能优于 ANN.

图 1 是唇高参数的实测值、支持向量回归预测值以及神经网络预测值的比较. 图 1(a)是实际跟踪的唇高变化曲线,图 1(b)是人工神经网络预测的结果,图 1(c)是支持向量回归预测的结果. 从图 1 可以看出,二者结果接近,但支持向量回归算法预测的结果较为平滑.

5 结束语

本文尝试用支持向量回归的方法进行唇动参数的预测,首先采用 K-L 变换有效地压缩了输入数据的维数,从而降低了模型的复杂度,然后采用交叉校

验和最速下降优化方法选择了最佳的支持向量回归参数. 实验结果表明它比传统的神经网络学习方法更为有效. 考虑到神经网络已经过多年的发展和完善,而支持向量回归则是一种正在发展的新的统计学习方法,还有更多的潜力可以挖掘,因此这种方法有很好的应用前景.



(a) 实侧值

(b) 神经网络预测值

(c) 支持向量回归预测值

图 1　唇高参数的实测值和预测值的比较

致谢　在 SVR 的训练过程我们应用了台湾国立大学 Chih-Chung Chang 和 Chih-Jen Lin 编写的 LIBSVM 的部分源代码[10],在此向他们表示衷心地感谢!

参 考 文 献

1　R R Rao, Tsuhan Chen. Cross-modal prediction in audio-visual communication. In: 1996 IEEE Int'l Conf on Acoustics, Speech, and Signal Processing (ICASSP'96). Piscataway, NJ, USA: IEEE, 1996. 2056~2059

2　Tsuhan Chen, R R Rao. Audio-visual interaction in multimedia communication. In: 1997 IEEE Int'l Conf on Acoustics, Speech, and Signal Processing (ICASSP'97). Piscataway, NJ, USA: IEEE, 1997. 179~182

3　R R Rao, Chen Tsuhan, Mersereau Russell M. Audio-to-visual conversion for multimedia communication. IEEE Trans on Industrial Electronics, 1998, 45(2): 15~22

4　T Chen. Audiovisual speech processing. IEEE Signal Processing Magazine, 2001, 18(1): 9~21

5　F Lavagetto. Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio-video synchronization. IEEE Trans on Circuits and Systems for Video Technology, 1997, 7(5): 786~800

6　Kyoung Ho Choi, Jenq-Neng Hwang. Baum-Welch hidden Markov model inversion for reliable audio-to-visual conversion. In: 1999 IEEE 3rd Workshop on Multimedia Signal Processing. Piscataway, NJ, USA: IEEE, 1999. 175~180

7　J J Williams, A K Katsaggelos, M A Randolph. A hidden Markov model based visual speech synthesizer. In: 2000 IEEE Int'l Conf on Acoustics, Speech, and Signal Processing (ICASSP'00). Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers Inc, 2000. 2393~2396

8　[美]V N Vapnik,张学工译. 统计学习理论的本质. 北京:清华大学出版社, 2000
(V N Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 2000)

9　Bernhard Schölkopf, Alex J Smola, Robert Williamson *et al*. New support vector algorithms. NeuroCOLT2, Tech Rep: NC2-TR-1998-031, 1998

10　Chih-Jen Lin. LIBSVM. Taibei: National Taiwan University, 2003. http://www.csie.ntu.edu.tw/~cjlin/

王志明　男,1968 年生,博士研究生,工程师,主要研究方向为语音可视化及汉语视位的建模.

蔡莲红　女,1945 年生,教授,博士生导师,主要研究方向为语音处理与合成、多媒体技术、生物特征识别等.

艾海舟　男,1964 年生,博士,副教授,主要研究方向为计算机视觉、模式识别.