

利用 SVM 的聚类算法在时间序列信号识别中的应用

汪永涛

(安徽财贸职业学院 电子信息系, 安徽 合肥 230601)

摘 要: 研究了一维时间序列信号识别的问题, 针对基于混合高斯模型的隐马尔科夫(HMM)编码准确率低的问题, 提出了一种利用多个支持向量机构造混合支持向量机, 从而为隐马尔科夫模型提供更精确的观测值编码和发生矩阵, 能有效的提高 HMM 在语音信号识别或者文字识别中的准确率. 本方法可以应用到语音识别, 文字识别以及生物信息处理等领域.

关键词: 隐马尔科夫模型; 支持向量机; 神经网络

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)03-0182-03

Using SVM and HMM for Serial Signal Processing

WANG Yong-tao

(Department of Electronic Information, Anhui Finance and Trade Vocational College, Hefei 230601, China)

Abstract: To aim the problem of one dimensional signal processing and recognition, such as speech recognition, we propose a mixture SVM model instead of mixture Gaussian model for Hidden Markov model. The method can be used in other areas such as biometrics.

Key words: HMM; SVM; neural network

1 引言

基于统计模型的机器学习方法近年来在人工智能领域得到了快速的发展. 支持向量机(SVM)是机器学习领域较为重视的一种识别分类模型.

虽然支持向量机在模式识别和机器学习领域得到了很好的研究和应用, 但缺点是: (1) 训练时间较长. 支持向量机的训练时间以训练样本个数的四次方递增, 因此对于大样本训练集合, 需要较长的训练时间, 使得其训练复杂度较高; (2) 仅针对两类问题. 通常的支持向量机模型的分割超平面是针对两类问题设计的, 对于多类问题, 需要进行多次两类问题比较, 从而使得多类问题的计算复杂度为类别个数的两次方, 这对于类别较多的问题, 会导致识别时间剧增, 效率急剧下降^[1].

文中设计了一种新颖的, 利用 SVM 对 HMM 进行编码的算法, 在保证计算效率的同时, 能有效的提高隐马尔科夫模型的识别准确率^[2-3].

2 隐马尔科夫模型

最早将隐马尔科夫模型(HMM)引入实际应用的是语音识别领域^[3]. HMM 描述了一个两层随机过程. 其中隐含层被认为是一个由有限个状态构成的随机状态机, 其可能包含的状态为: $S = \{s_1, \dots, s_K\}$. 在任意时刻 t , 隐含层所在的状态只由其前一刻 $t-1$ 所在系统状态决定: $P(S_t | S_1, \dots, S_{t-1}) = P(S_t | S_{t-1})$; 即马尔科夫性. 该隐含层在各个状态之间以一定的概率进行转移, 构成一个概率转移矩阵:

$$A = \{a_{ij} | a_{ij} = P(S_t = j | S_{t-1} = i)\} \quad (1)$$

该公式表示当时刻 $t-1$ 时系统处于状态 i , 则在时刻 t 时系统处于状态 j 的概率. 隐含层的状态决定了系统的特性.

但是最为外部观测者, 并不能直接观测到隐含层的变换, 而只能观测到有隐含层产生的一组观测值, 任意时刻 t 所能观测到的值只与当前时刻的隐含层状态有关: $P(O_t | O_1, \dots, O_{t-1}, S_1, \dots, S_t) = P$

$(O_t | S_t)$.

由此 HMM 构成了一个由隐含层到观测层的发生矩阵:

$$B = \{b_{jk} | b_{jk} = P(O_t = o_k | S_t = j)\} \quad (2)$$

该公式表明,如果在时刻 t ,系统隐含状态为 j 时,系统外部观测为 o_k 的概率.

在对一个随机过程信号进行识别的过程中,需要寻找到一个合适的隐马尔科夫模型 M^* ,使得在给定的观测信号序列 X 时,使得如下公式得到满足:

$$M^* = \max_M P(M | X) \quad (3)$$

由于无法直接计算以上后验概率,通常利用 Bayes 法则将后验概率模型转换成先验概率与似然函数的乘积:

$$M^* = \max_M p(M | X) \propto p(X | M) p(M) \quad (4)$$

在通常的 HMM 训练过程中,对于发生矩阵 B 的求解通常利用混合高斯模型,预先将所有观测特征值进行聚类 and 编码操作,然后利用 EM 算法进行最优化求解.但是常规的混合高斯模型利用基于极大似然的方法对观测值进行编码,会导致所预估的聚类方差和均值与实际数据发生较大差距.

因此,利用神经网络或者 SVM 等更精确的数学模型,并且经过训练来实现对观测值的分类和编码,能有效的提高对观测数据的分类准确率.文中采用支持向量机模型来对观测值进行聚类 and 编码,从而提高 HMM 的识别准确率.

3 基于 SVM 的混合模型

支持向量机(SVM)是一个线性分类模型,基本原理是以样本中间那些最难以区分的样本作为支持向量,从而构造一个分割曲面. SVM 的最初版本是针对线性分类问题,但是可以利用一个核函数将非线性的低维数据从低维空间 R^n 投影到一个高维空间 F 中,从而使得原来不可分的样本在高维空间中能线性可分.在实际应用中,最常用的核函数是^[4]:

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \quad (5)$$

对于一组给定的训练样本以及对应的类别标志 $\{x_i, y_i\}$,其中 $x_i \in R^n$, $y_i \in \{-1, +1\}$,如果选定核函数 $k(x, y)$ 以及 SVM 相关的参数 C ,则通过计算如下公式获得一个最优的分割超平面:

$$W(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (6)$$

并且使得: $\sum_{i=1}^N \alpha_i y_i = 0$ 和 $0 \leq \alpha_i \leq C$. 最终得到的一个分割面可用如下公式表示:

$$s(x) = \sum_{i=1}^N \alpha_i y_i k(x_i, x) + b \quad (7)$$

如果 $s(x) > 0$ 则意味这 x 归类为正,否则归类为负.

通常利用分类器来建立混合模型的方法是建立多个分类器,然后做一对一比较,从而获知最优的类别达到聚类或者编码的目的.

文中使用一种并发的支持向量机方法来实现聚类.

假设有 M 个聚类并对应着 M 个支持向量机分类器,则对于一个输入测试样本,其输出表示为:

$$f(x) = \sum_{m=1}^M w_m(x) s_m(x) \quad (8)$$

其中 $s_m(x)$ 为第 m 个 SVM 分类器的输出, $w_m(x)$ 为该分类器的权重. $w_m(x)$ 的具体值可以根据神经网络经过 BP 算法训练得到,其数学模型可表述为计算如下公式的最小值:

$$\sum_{i=1}^N [f(x_i) - y_i]^2 \quad (9)$$

混合模型的具体构建过程如下:

(1) 首先将原始训练数据分割成互补不叠的若干个子集,每个自己包含的样本个数为: N/M ;

(2) 对每个子集训练一个支持向量机分类器,在后续处理中,将固定使用这组 SVM,而其对应的权重 $w_m(x)$ 则依据公式(9)计算得到;

(3) 在后续的训练过程中,对任意一个输入样本及其分类标志 $\{x_i, y_i\}$ 将被指定给一个具体的支持向量机,其原则是先将 SVM 按照权重排序,然后将该训练样本赋给第一个样本数不足 $N/M+1$ 的支持向量机,然后该向量机在新添加的样本上重新训练.该过程可持续多次,直到所有新加的样本训练完成为止.

4 SVM 模型的后验概率

由于公式(8)所输出的是经过权重处理的类表标志,并不能直接用于计算一个样本的后验概率,因此,必须重新计算混合 SVM 的输出,用以表示后验概率.文中采用的方法是利用类似 SVM 进行线性规划所采用的拟合方法,其公式表达为:

$$\begin{aligned} p(y = +1 | x) &= g(f(x), A, B) \\ &= \frac{1}{1 + \exp(Af(x) + B)} \end{aligned} \quad (10)$$

其中的 A 和 B 可以通过极大似然估计来获得, 其具体方法如文献[5]中所述, 首先构造一个训练集合 (p_i, t_i) , 其中 $p_i = g(f(x), A, B)$, $t_i = (y_i + 1)/2$. 可以通过计算下面公式的最小信息熵来获得最优的 A 和 B 值:

$$-\sum_i [t_i \log(p_i) + (1 - t_i) \log(1 - p_i)] \quad (11)$$

由于单一的 SVM 是针对两类问题设计的分类器, 所以对于多类问题, 需要训练多个混合支持向量机来进行筛选. 最终对于一个测试样本, 将比对任意两个混合 SVM 的输出值, 并挑选出后验概率最大的作为其类别结果.

最终, 需要将一个样本的后验概率 $p(q|x)$ 转换为 HMM 模型的发生概率 $p(x|q)$, 其依据是贝叶斯公式:

$$p(x|q) \propto \frac{p(q|x)}{p(q)} \quad (12)$$

5 实验结果

实验对比了采用本文混合 SVM 的 HMM 识别准确率与常规的基于混合高斯模型的 HMM 的识别准确率, 如表 1 所示.

表 1 实验结果对比

实验方法	识别准确率
混合高斯+HMM	87.4%
混合 SVM+HMM	93.4%

从表 1 中可以看到, 利用本文方法的语音信号识别率要高于传统的基于混合高斯模型的 HMM 识别率.

为了测试本文算法对其他类型的一维序列信号的试验效果, 对一组英文文字图像进行训练和识别. 一共采集了 200 页, 近 10000 个手写英文单词进行训练(如图 1 所示). 在训练时, 首先将一行文字截取出来, 然后以每 25 个像素为单位, 每 10 个像素重叠为基准, 对一行文字进行采用, 并提取其特征. 然后分别利用混合高斯模型和混合 SVM 模型分别对观测到的样本进行聚类 and 编码, 然后利用 HMM 进行分类和识别.

This article is on palm mobilts. Its as wed is these

图 1 英文文字图样

最终的文字识别率分别为 76.1%(混合高斯)和 85.2%(混合 SVM). 由此可以得出结论, 使用本文方法能有效的提高 HMM 的编码准确率, 从而进一步得到更高的系统识别率.

6 结束语

在本文中, 提出了一种基于混合支持向量机模型的一维信号识别模型. 该模型通过利用多个加权的支持向量机构成一个更为精确的支持向量机模型, 并能直接计算出一个测试样本的后验概率. 通过将原始训练集合划分成若干个子集, 训练得到一组混合支持向量机, 由此可以计算出任意测试样本所属的类别, 从而实现了观测样本的聚类 and 编码. 在此基础上进行的 HMM 训练和识别能获得更高的识别准确率.

参考文献:

- [1] 高清伦, 谭月辉, 王嘉祯. 基于离散隐马尔科夫模型的语音识别技术[J]. 河北省科学院学报, 2007, 3(3): 110-114.
- [2] Golowich S E, Sun D X. A support vector/hidden markov model approach to phoneme [J]. Proceedings of the Statistical Computing Section, 1998, 42(3): 125-130.
- [3] 欧智坚, 王作英. 从线性预测 HMM 到一种新的语音识别的混合模型 [J]. 电子学报, 2002, 11(9): 433-436.
- [4] 张浩然, 韩正之, 李昌刚. 基于神经网络的语音合成环节中英语单词音节划分的技术[J]. 计算机工程与应用, 2003, 14(9): 435-439.
- [5] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods [J]. Advances in LargeMargin Classifiers, MIT Press, 1999, 61(4): 636-660.
- [6] S E Krüger, Schafföner M, Katz M, et al. Speech recognition with support vector machines in a hybrid system [C] // IPROC INTERSPEECH-2005. [s. l.]: IEEE, 2005: 993-996.
- [7] 颜文胜. 无线传感器网络节点部署算法的优化研究[J]. 计算机仿真, 2011, 28(4): 126-129.

作者简介:

汪永涛 男, (1968—), 硕士, 讲师. 研究方向为计算机网络路由、计算机仿真、神经网络、计算机应用、自动化控制、计算机检测.