在变长模式识别中利用测地距离的非线性插值进行特征选取*)

黄石磊 谢 湘 匡镜明
(北京理工大学信息科学技术学院电子工程系 北京 100081)

摘 要 讨论了变长模式识别中的特征选择问题。采用基于测地距离(Geodesic Distance)的非线性插值来进行特征选择,使得变长的模式映射为等长的模式,从而可以使用传统的等长模式的方法来解决变长模式识别问题。用非特定说话人的汉语孤立词识别来验证提出方法的性能,并采用支持向量机(Support Vector Machine, SVM)作为基本的分类方法。实验结果表明,提出的方法可以获得比传统方法诸如线性插值更好的性能,而计算量仅有很少增加。

关键词 模式识别,测地距离,支持向量机,特征选择

Feature Selection in Length-variant Pattern Recognition Using Non-linear Interpolation Based on Geodesic Distance

HUANG Shi-Lei XIE Xiang KUANG Jing-Ming
(Department of Electronic Engineering, Beijing Institute of Technology, Beijing 100081)

Abstract Discuss the feature selection in length-variant pattern recognition. Non-linear interpolation based on geodesic distances is used in feature selection. Length-variant patterns are mapped into Length-invariant patterns by a non-linear interpolation; then traditional pattern recognition methods for length-invariant patterns can be used in solving the recognition problem. Experiments of speaker-independent Mandarin isolated word recognition were performed to evaluate the performance of the proposed method. And support vector machine (SVM) is used as basic classification method. Experimental result shows that the proposed method has achieved better performance than traditional method such as linear interpolation, and computational complexity increased slightly.

Keywords Pattern recognition, Geodesic distance, Support vector machine, Feature selection

1 引言

在模式识别中,某些模式是一种变长的模式,典型的如语音信号。在利用支持向量机(Support vector machine, SVM)[1]等静态模式识别方法进行分类时,需要把这些变长的模式映射为等长的模式。因此,如何从变长模式的特征中选择具有"代表性"的特征子集是一个十分重要的问题。

在典型的语音识别问题中,从一个语音样本可以分帧提取基本的声学特征,这样可以得到一个矢量序列,作为这个语音样本的特征。注意到真实的发音的长度是变化的,此特征矢量序列的长度也是变化的。已有一些方法可以将不同长度的特征矢量序列映射为等长的矢量序列,进而合并为一个矢量作为诸如SVM等分类器的输入矢量[2]。

2000年,Tenenbaum提出用测地距离(Geodesic Distance)进行特征的降维[3]。其基本思想是在流型(Manifold)中相邻点对的欧氏距离是测地距离的较好近似,相距较远的点对的测地距离,使用邻近点距离的一条最短路径来估计。基于测地距离的非线性变换在数据挖掘、图像处理、人脸识别等方面得到了广泛的应用。

在本文中,利用基于测地距离的插值将变长模式映射为等长的模式。将矢量序列中时序上相邻的矢量看成测地距离上相邻点,而把矢量序列看成一个测地距离路径。经过插值以后,不同长度的矢量序列变为等长,并可采用静态的模式方法进行处理。在本文中使用SVM作为静态模式识别器。

2 基本识别系统

尽管HMM、TDNN等一些方法可以处理变长模式问题[4],但是运算量巨大,在某些应用场合,诸如手持设备上并不一定适用。而一些静态模式识别方法,运算简单,也能达到较好的分类/识别效果[5]。

相对于基本的静态模式识别系统,处理不等长模式的识别问题一般首先从样本中提取一定的特征序列,然后对这个特征序列进行一定的处理,使之成为等长的模式,最后传递给分类器[6]。整个过程如图1所示。

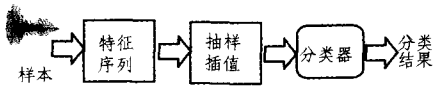

图1 不等长模式处理过程

假定从某个样本中提取了特征矢量序列(v_1, v_2,…,v_k),不同的样本具有不同的特征矢量的个数,简单的可以采用线性插值的方法,得到一个具有固定矢量个数的序列(u_1,u_2,…,u_L)。把这些矢量合并到一起,形成一个大的矢量:

$$x=(u_1^T,u_2^T,\cdots,u_L^T) \tag{1}$$

用这个矢量来表示相应的样本。根据识别系统采用的识别器,例如SVM,这个矢量将用来训练模板或者用于分类。

3 基于测地(Geodesic)距离的非线性插值

假设包含k个矢量的矢量序列(v_1, v_2,…,v_k),它们形成

*)自然科学基金(No. 60372089)。黄石磊 博士生;匡镜明 教授。

一条路径，或者称为测地距离路径。我们可以定义矢量序列中两个矢量(v_i, v_j)的测地距离为

$$D_{geo}(v_i,v_j)=\begin{cases}0, & \text{if } i=j\\ \sum_{n=i}^{j-1}D_{Euc}(v_n,v_{n+1}), & \text{if } j>i\\ D_{geo}(v_j,v_i), & \text{if } j<i\end{cases} \quad (2)$$

其中 D_{Euc} 是两个矢量的欧氏距离：

$$D_{Euc}(x,y)=\sqrt{\|x-y\|} \quad (3)$$

假定用 L 个矢量的矢量序列($u_1,u_2,\cdots,u_L$)来代表上述 k 个矢量的矢量序列。其中的矢量 u_j 是测地距离路径(v_1, $v_2,\cdots,v_k$)上的点。也就是说，u_j 是矢量(v_1, $v_2,\cdots,v_k$)中的元素或者是线段(v_i, v_{i+1})上的点，并且

$$u_1=v_1,u_L=v_k$$
$$D_{geo}(u_i,u_{i+1})=D_{geo}(u_j,u_{j+1}),i\neq j,i,j=1,2,\cdots,L-1 \quad (4)$$

一般地，矢量序列($u_1,u_2,\cdots,u_L$)的数量比矢量序列(v_1, $v_2,\cdots,v_k$)要少。如果 u_i 在线段 (v_{p-1}, v_p) 上，且 u_j 在线段(v_q, v_{q+1})上，则(u_i, $v_p,\cdots$, v_q u_j,) 构成了一个测地距离路径：

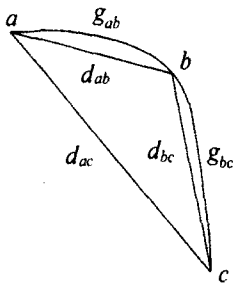

图 2　测地距离和欧氏距离的比较

g_{ab}/g_{bc} 为测地距离，$d_{ab}/d_{bc}/d_{ac}$ 为欧氏距离。

$$D_{geo}(u_i,u_j)=D_{geo}(u_i,v_p)+D_{geo}(v_p,v_q)+D_{geo}(v_q,u_j)=$$
$$D_{Euc}(u_i,v_p)+D_{geo}((v_p,v_q)+D_{Euc}(v_q,u_j) \quad (5)$$

在理想情况下，如果两个特征矢量的序列 V_1 和 V_2 具有相同的测地距离路径，其相应的“代表”矢量序列分别是 U_1 和 U_2，并考虑如下的情况：

$$V_1=(v_1^1,v_2^1,\cdots,v_i^1,v_{i+1}^1,\cdots,v_k^1)$$
$$V_2=(v_1^2,v_2^2,\cdots,v_i^2,v_{i+1}^2,v_{i+2}^2,\cdots,v_{k+1}^2) \quad (6)$$

其中 $v_j^1=v_j^2$,($j=1,2,\cdots,i$)；$v_j^1=v_{j+1}^2$,($j=i+1,\cdots,K$)；并且 v_{i+1}^2 在线段(v_i^1,v_{i+1}^1)上，于是有

$$D_{geo}(v_i^1,v_{i+1}^1)=D_{geo}(v_i^2,v_{i+2}^2)$$
$$=D_{geo}(v_i^2,v_{i+1}^2)+D_{geo}(v_{i+1}^2,v_{i+2}^2) \quad (7)$$

这就意味着在计算 V_2 相应的矢量序列 U_2 的时候 v_{i+1}^2 将被忽略，并且 U_1 和 U_2 将完全相同。这表明，如果 V_1 和 V_2 在空间的轨迹相同，那么计算出来的代表特征矢量序列 U_1 和 U_2 也相同，而与在轨迹上某些点的密集程度无关。如果模式之间的区别仅仅在于空间的轨迹，那么采用本文的方法计算代表特征矢量序列的时候就十分有效。

4　具体实现

如何从已有的 k 个矢量序列(v_1, $v_2,\cdots,v_k$)中求 L 个矢量序列($u_1,u_2,\cdots,u_L$)，使之满足式(4)，可以采用如下的算法。

(a)计算原始矢量序列的测地距离矩阵：

$$d_{ij}=D_{geo}(v_i,v_j),i,j=1,\cdots,k \quad (8)$$

以及每个矢量到第一个矢量的测地距离：

$$d_i=D_{geo}(v_1,v_i),i=1,\cdots,k \quad (9)$$

由式(4)可知，($u_1,u_2,\cdots,u_L$)中各相邻点的测地距离是相等的，因此可以得到 u_i 和 u_{i+1} 之间的测地距离为

$$d_\Delta=D_{geo}(u_i,u_{i+1})=d_k/(L-1) \quad (10)$$

(b)初始化，设 $u_1=v_1$。

(c)从已计算出的 $u_1,\cdots,u_i$ 中求 u_{i+1}。

由于 $D_{geo}(u_i,v_j)$ 随 j 增长而变大，于是可以找到 v_j 和 v_{j+1}，使得 $D_{geo}(u_i,v_j)\leqslant d_\Delta$，并且 $D_{geo}(u_i,v_{j+1})\geqslant d_\Delta$，从而可以确定 u_{i+1} 是在(v_i, v_j)之间的线段上，于是有

$$\begin{cases}u_{i+1}=\lambda v_i+(1-\lambda)v_{i+1}\\ D_{Euc}(u_{i+1},v_i)=d_\Delta-D_{geo}(u_i,v_j)\end{cases} \quad (11)$$

其中 $\lambda\in[0,1]$，为实数。

根据以上方程可以求得 λ 和 u_{i+1}。

(d) 重复步骤(c)，直到计算得 u_{L-1}。

(e)$u_L=v_k$。

经过以上步骤，可由(v_1, $v_2,\cdots,v_k$)求得($u_1,u_2,\cdots,u_L$)，之后把 L 个矢量合并到一起，形成一个大的矢量，这个矢量将作为 SVM 的输入矢量 x。

5　支持向量机

支持向量机(Support Vector Machine)是一种新兴的、基于结构风险最小化的模式分类技术[1]，它在各种应用中都取得了较好的效果[7]。

设(X_i,Y_i)，$1\leqslant i\leqslant N$，$X_i\in R^d$，$Y_i\in\{-1,1\}$ 表示一个训练样本集，d 为样本特征空间的维数，Y_i 为样本 X_i 的所属类别。若有超平面 $W\cdot X+b=0$ 正确地将样本划分成两类，则称该样本集是线性可分的。而最优的分类超平面使两类样本到超平面的最小距离为最大。设对所有样本 X_i，使 $|W\cdot X+b|$ 的最小值为 1，则样本与此最优分类超平面的最小距离为 $1/\|W\|$。最优分类超平面应满足约束：

$$Y_i(W\cdot X+b)\geqslant 1,i=1,\cdots,N \quad (12)$$

其中，W 和 b 的优化条件应是使两类样本到超平面最小的距离之和 $2/\|W\|$ 为最大，即最小化 $W\cdot W/2$。

上述最优分类面问题可转化为其对偶 Lagrange 问题，并求解，得

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jY_iY_j(X_i,X_j) \quad (13)$$

其中 α_i 为与每个样本对应的 Lagrange 乘子。可以证明，其中只有一部分 α_i 不为零，对应的样本就是支持向量。解上述问题后，得到最优分类函数：

$$f(X)=\mathrm{sgn}(\sum_{i=1}^{N}\alpha_i^*Y_i(X_i\cdot X)+b^*) \quad (14)$$

对于线性情况下不可分的情形，可以通过一个非线性映射 $\Phi(x)$，把输入的数据从原特征空间映射到一个高维的特征空间 Ω，再在高维的特征空间中建立最优的分类超平面。在线性情况只用到了原空间的点积运算，在非线性空间也只考虑在高维特征空间 Ω 的点积运算 $\Omega(x)\cdot\Omega(y)=K(x,y)$，不必明确知道 $\Omega(x)$。$K(x,y)$ 称为核函数，核函数的选取应使其为特征空间的一个点积，即存在函数 $\Phi(x,y)$，使 $\Phi(x)\cdot\Phi(y)=K(x,y)$。在非线性情况下，支持向量机对分类问题成为最大化函数(15)，相应的分类函数变为(16)：

$$Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jY_iY_jK(X_i\cdot X_j) \quad (15)$$

$$f(X)=\mathrm{sgn}(\sum_{i=1}^{N}\alpha_i^* Y_i K(X_i \cdot X)+b^*) \quad (16)$$

6 实验结果

为了比较采用测地距离的非线性插值和一般的线性插值方法对于不等长模式识别的性能，在非特定人汉语孤立词(命令)识别系统中分别采用这两种方法，得到不同的实验结果。

本文中的汉语孤立词识别系统的框图如图 3 所示。实验中所采用的数据库共包含 60 个说话人(30 男和 30 女)的语音。其中一半的语音(15 男和 15 女)作为训练集，其余的部分作为测试集。在训练集中，每个说话人分别重复 50 个汉语命令词 4 遍，这些命令词都是手持设备的一些控制命令，命令词的长度为 2 到 4 个音节。在测试集中，每个说话人对和训练集相同的 50 个命令词重复 4 遍。

数字信号形式的语音波形首先经过预处理，包括 VAD、预加重和分帧。每 10ms 一帧，帧长为 25ms，然后从每帧语音中提取语音的初始声学特征向量。本文中的声学特征使用 12 阶(包括 0 阶)MFCC 和对数能量，共 13 维。

假定从语音波形中提取了 k 个声学特征向量(v_1，v_2，…，v_k)。参照公式(1)，采用线性和非线性的插值得到 L 个代表性的特征向量。在本文的实验中，经过线性插值和抽取，得到 8～16 个具有“代表性”的声学特征向量来代表一条语音。这 8～16 个向量组合起来的高维向量作为 SVM 识别器的输入向量。

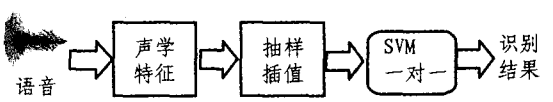

图 3 基于 SVM 的汉语孤立词识别系统的框图

在训练和测试过程中，这个固定大小的向量 x 都是直接提供给 SVM 分类器。本实验中的孤立词识别问题是多分类问题，实验中采用的是一对一的方式构造 SVM 识别器[7]。对于 50 个命令中的任意两个数字，都构造一个单独的 SVM 分类器，因此总共有 1225 个单独的 SVM 分类器构成总的识别器，这个总的识别的最终结果是由各个单独识别的识别结果投票得到的。这种情况下，虽然在每次识别一个样本时需要较多次判断，但是因为 SVM 的判决函数式(16)较为简单，其总运算量依然在可接受的范围内。

实验中，SVM 工具包采用 Chang Chih-Chung 和 Lin Chin-Jen 的 libSVM[8]。SVM 的核函数选用了线性核函数和径向基(RBF)函数：

$$k(x,y)=x \cdot y \quad (17)$$

$$k(x,y)=\exp(-|g| \cdot \|x-y\|^2) \quad (18)$$

其中 g 是常数。

实验结果如表 1 和表 2 所示。

表 1 使用线性核函数时不同插值方法的词错误率(%)

代表矢量的个数	线性插值	非线性插值	词错误率下降
8	10.07	10.03	0.40
12	9.50	8.95	6.15
16	9.18	8.43	8.90
20	9.23	8.58	7.58

表 2 使用 RBF 核函数时不同插值方法的词错误率(%)

代表矢量的个数	线性插值	非线性插值	词错误率下降
8	9.80	9.53	2.83
12	8.77	8.05	8.94
16	8.47	7.60	11.45
20	8.53	7.70	10.78

7 分析

从实验结果来看，采用本文中的非线性插值方法得到的词错误率(Word Error Rage，WER)明显下降。比如在使用 16 个代表性的特征矢量时，在 RBF 核函数条件下，WER 从 8.47%下降到了 7.60%，相当于下降了大约 11.45%。

同样可以看到，从原始矢量序列中选择的矢量越多，识别的正确率越大。但是，多到一定程度后，并没有进一步提高，甚至略微有所下降。矢量越多，相应的信息量越大，不同的模式(词)之间的区分就越明显。但是对于 SVM 来说，特别是采用非线性核函数，矢量的数量越多，充分训练所需的样本也越来越多。但是实际样本的总数有限，所以到了一定的程度之后，训练不充分了，性能出现了下降。同时，选择代表性的矢量越少，非线性方法和线性插值所得到的性能越接近。这是由于矢量过少，两种方法之间的区别也变小了。极限情况下，如果只选择两个代表性的矢量，实际上两种方法求得的矢量是一样的，性能也应该完全相同。

比较不同核函数之间的性能，RBF 核函数和线性函数相比具有更低的 WER，并且采用非线性插值方法获得代表性的特征矢量的时候，其 WER 的相对下降也更多。

总的来说，采用本文中提出的基于测定距离的非线性插值的方法增加了特征矢量求取时的计算量，但是相对于 HMM 等更复杂的方法，这种方法的运算量还是要小很多。

结论 本文提出采用基于测地距离的非线性插值方法用于非等长模式的特征选择问题中，并在基于 SVM 的语音识别应用中对比了非线性插值方法与传统线性插值方法的性能。非特定人孤立词语音识别实验的结果表明，尽管计算量有一定的增加，采用提出方法取得了更好的识别效果。进一步研究中需要考虑语音的驻留时间分布对特征选择的影响。

参考文献

1 Vapnick V. The Nature of Statistical Learning Theory New York: Springer-Verlag, 1995

2 Smith N, Gales M. Speech Recognition using SVMs. In: Proceedings of the NIPS 2001, 2001. 1197～1204

3 Tenenbaum J B, et al. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500): 2319～2323

4 Wilpon J G, Lee C H, Rabiner L R. Application of Hidden Markov Models for Recognition of a Limited Set of Words in Unconstrained Speech. ICASSP-89, vol 3. 254～257

5 Ganapathiraju A, Hamaker J, Picone J. Applications of Support Vector Machines to Speech Recognition. IEEE Transactions Signal Processing, 2004, 52(8): 2348～2355

6 Bernal-Chaves, et al. Multiclass SVM-based isolated-digit recognition using a HMM-guided segmentation. NOLISP-2005, 2005. 137～144

7 Cones C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20(3): 273～297

8 Chang Chih-Chung, Lin Chih-Jen. LIBSVM: A Library for Support Vector, Machines. Depart. of Computer Science and Information Engineering, National Taiwan University. Software available at http://www.csie.ntu.edu.tw/～cjlin/libsvm/. 2004