一种新的子空间模式识别方法

王 洪，吕幼新，向敬成
(电子科技大学电子工程学院 成都 610054)

【摘要】提出了一种改进的LSM-ALSM子空间模式识别方法，将LSM的旋转策略引入ALSM，使子空间之间互不关联的情况得到改善，提高了ALSM对相似样本的区分能力。讨论中以性能函数代替经验函数来确定拒识规则的参数，实现了识别率、误识率与拒识率之间的最佳平衡；通过对有限字符集的实验结果表明，LSM-ALSM算法有效地改善了分类器的识别率和可靠性。

关 键 词 学习子空间； 性能函数； 散布矩阵； 最小描述长度
中图分类号 TN919.8 TP391.4; 文献标识码 A

A New Subspace Pattern Recognition Method

Wang Hong，Lü Youxin，Xiang Jingcheng
(School of Electronic Engineering, UEST of China Chengdu 610054)

Abstract A new subspace algorithm(LSM-ALSM) is discussed in this paper. By rotating the subspaces of ALSM, LSM changes their independent relations and improves the ALSM classifier's recognition capacity in alike characters. To realize the best balance between recognized rate, rejected rate and error rate, we presents performance function instead of experience function to determine the refuse rules. The experiment with limited character set shows that LSM-ALSM method improves the recognized rate and credibility effectively.

Key words study subspace; performance function; spread matrix; minimum description length

在子空间模式识别方法中，一个线性子空间代表一个模式类别，该子空间由反映类别本质的一组特征矢量张成，分类器根据输入样本在各子空间上的投影长度将其归为相应的类别。典型的子空间算法有以下三种[1, 2]：CLAFIC(Class-feature Information Compression)算法以相关矩阵的部分特征向量来构造子空间，实现了特征信息的压缩，但对样本的利用为一次性，不能根据分类结果进行调整和学习，对样本信息的利用不充分；学习子空间方法(Leaning Subspace Method, LSM)通过旋转子空间来拉大样本所属类别与最近邻类别的距离，以此提高分类能力，但对样本的训练顺序敏感，同一样本训练的顺序不同对子空间构造的影响就不同；平均学习子空间算法(Averaged Learning Subspace Method, ALSM)是在迭代训练过程中，用错误分类的样本去调整散布矩阵，训练结果与样本输入顺序无关，所有样本平均参与训练，其不足之处是各模式的子空间之间相互独立。针对以上问题，本文提出一种改进的子空间模式识别方法。

1 子空间模式识别的基本原理

1.1 子空间的分类规则

子空间模式识别方法的每一类别由一个子空间表示，子空间分类器的基本分类规则是按矢量 x 在各子

收稿日期：2003 – 10 – 14
作者简介：王 洪(1974 –)，男，在职博士生，助教，主要从事模式识别、信号处理方面的研究.

空间上的投影长度大小，将样本归类到最大长度所对应的类别，$\boldsymbol{x}$ 在类 $w^{(i)}$ 的子空间上投影长度的平方为

$$g(\boldsymbol{x})=\underset{j=1,2,\cdots,p}{\operatorname{argmax}}\sum_{k=1}^{M^{(j)}}(\boldsymbol{b}_k^T\boldsymbol{x})^2 \tag{1}$$

式中 函数 $g(x)$ 称为分类函数；$\boldsymbol{b}_k$ 为子空间基矢量。两类的分类情况如图1所示。

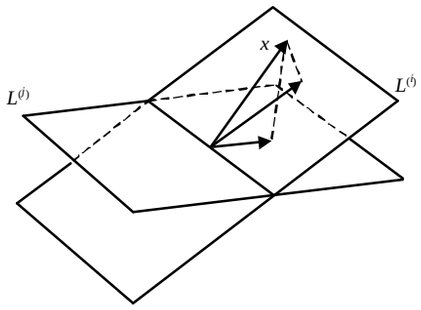

图1 矢量分类示例

1.2 拒识规则

当满足下列条件之一时子空间分类器拒识

$$\underset{i}{\operatorname{argmax}}[(\boldsymbol{x}^T\boldsymbol{P}^{(i)}\boldsymbol{x})^{1/2}]<m_1\|\boldsymbol{x}\| \tag{2}$$

$$\left|(\boldsymbol{x}^T\boldsymbol{P}^{(i)}\boldsymbol{x})^{1/2}-(\boldsymbol{x}^T\boldsymbol{P}^{(j)}\boldsymbol{x})^{1/2}\right|<m_2\|\boldsymbol{x}\| \tag{3}$$

式中 i 和 j 是使 $\boldsymbol{x}$ 投影长度最大和次最大的两个类别的子空间；m_1 和 m_2 是介于[0,1]之间的系数。当输入矢量 $\boldsymbol{x}$ 在各子空间上的投影的最大长度小于其自身长度的一个比例量(如85%)时，予以拒识，或投影长度最大与次最大的差值小于矢量自身长度的一个比例(如10%)时，予以拒识。

1.3 子空间的构造方法

CLAFIC的构造方法是用训练样本特征的散布矩阵去估计样本的相关矩阵，根据特征值的大小选取部分特征向量作为各类别子空间的基矢量。LSM和ALSM的构造方法都是建立在CLAFIC基础上，LSM是找出样本所属类别和最近邻类别，旋转对应的两个子空间，使样本在两个子空间上的投影距离拉大，其旋转方法为

$$\begin{cases}L^{(p)'}=(\boldsymbol{I}+r_p\boldsymbol{x}\boldsymbol{x}^T)L^{(p)}\\ L^{(q)'}=(\boldsymbol{I}-r_q\boldsymbol{x}\boldsymbol{x}^T)L^{(q)}\end{cases} \tag{4}$$

式中 p 表示正确类别；q 表示被误分的类别；r_p 和 r_q 为学习率。文献[2]提出的平均学习子空间方法可以看作是一种迭代的CLAFIC算法，其算法步骤如下：

1) 计算各类的初始散布矩阵 $\boldsymbol{S}^{(j)}(0)=\sum_{i=1}^{n^{(j)}}\boldsymbol{x}_i^{(j)}\boldsymbol{x}_i^{(j)T}$，$\boldsymbol{x}_i^{(j)}$ 表示第 j 类模式的第 i 个样本矢量；

2) 同CLAFIC方法一样，用 $\boldsymbol{S}^{(j)}(t)$ 中特征值较大的特征矢量形成子空间基矢量；

3) 将所有的训练样本进行分类，找出每个类别的两类错误样本集，对散布矩阵进行调整

$$S^{(j)}(t+1)=S^{(j)}(t)+a\sum_{x_i^{(j)}\in A}\boldsymbol{x}^{(j)}\boldsymbol{x}^{(j)^T}-b\sum_{x_i^{(k)}\in B}\boldsymbol{x}^{(k)}\boldsymbol{x}^{(k)^T} \tag{5}$$

式中 a 和 b 是学习率；$x_i^{(j)}$ 是属于类 $\boldsymbol{w}^{(j)}$ 但被错误地分到其他类别的样本；$x_i^{(k)}$ 是属于其他类但被错误分到 $w^{(j)}$ 的样本；

4) 返回步骤2)继续训练，当达到指定的迭代次数或样本的分类精度开始下降为止。

1.4 子空间分类方法的讨论

子空间模式识别的关键环节是分类规则、拒识规则、子空间构造和子空间维数，其中子空间构造方法尤为重要。CLAFIC算法提供了一种空间映射的方法，但由于没有学习调整作用而一般不直接采用。LSM对样本的训练次序敏感，其识别精度不如ALSM算法[3]，而ALSM算法的识别精度虽然较高，但还存在一定问题。

首先，在ALSM识别器中增强拒识规则，识别率下降了可靠性增加却不明显，没有起到此消彼涨的作用。其次，子空间构造方法保证了基矢量的正交性，但子空间之间却不是正交的。另外，从原理上看，ALSM算法的实质是通过迭代来调整散布矩阵，由于散布矩阵是所有训练样本的特征矢量转置相乘之和，错误样本相对于训练样本的总数来说很少，因此这种调整的作用有限。此外，误分类的样本中相当一部分是由于相似类别的样本造成，这些样本的区别是局部的，ALSM算法只能进行全局调整，为了少量局部特性去反复调整表达总体特性的散布矩阵不合理。

针对上述问题，本文在下面的讨论中将LSM和ALSM相结合，发挥各自的优势：ALSM算法在建立类别

的子空间上比较有效，但各类别的子空间之间互不相关，LSM算法通过旋转子空间来增大子空间之间的夹角，既对相似样本进行针对性调整，也改变了各子空间之间不相关联的情况；LSM算法的特点是局部调整，ALSM算法的特点是全局平均，两种算法结合体现了互补作用。

2　改进的LSM-ALSM算法

2.1　子空间的构造

ALSM用两类错误分类样本来调整散布矩阵，调整后再建立新的子空间。本文则用ALSM建立的子空间对训练样本分类，用错误分类的样本去调整相应类别及其最近邻类别的子空间。

2.2　改进分类规则

为了加大类别间的区分度，不妨对投影长度再乘上一个加权系数。ALSM子空间基矢量的选择是取决于该矢量对应的特征值大小，即特征空间被变换后的信息有效性程度是靠特征值大小来衡量，本文也采用特征值来加权投影长度，改进后的分类规则为[3]

$$g(\boldsymbol{x})=\underset{j=1,2,\cdots,p}{\operatorname{argmax}}\sum_{k=1}^{M^{(j)}}\frac{l_{kj}}{l_{1j}}(\boldsymbol{b}_k^T\boldsymbol{x})^2 \tag{6}$$

式中　l_{kj}为子空间基矢量对应的特征值，按从大到小排序。

2.3　改进拒识规则

拒识规则是排除误识的一道有效屏障，也是分类规则的延伸和补充。在式(2)、(3)中，m_1和m_2取的是经验值，由于识别对象、样本特征、子空间维数等都会影响投影长度的差异程度，因此这种经验方法不可靠。实际应用识别系统应符合“宁拒不误”的原则，但拒识率的增加会导致识别率的降低，拒识规则的确定也就是追求识别、拒识和误识之间的一个最佳折衷[4]。以性能函数PF确定m_1和m_2最佳取值的方法为

$$PF(m_1,m_2)=C_cR_c(m_1,m_2)-C_wR_w(m_1,m_2)-C_rR_r(m_1,m_2) \tag{7}$$

式中　C_c、C_w和C_r分别为正确识别获益、错误识别代价和拒绝识别代价；$R_c(m_1,m_2)$、$R_w(m_1,m_2)$、$R_r(m_1,m_2)$分别为以m_1和m_2为参数的识别率、误识率和拒识率。这里首先确定m_1和m_2的范围，在这个范围中以0.01为步长，通过循环求出最佳的系数。

2.4　子空间的维数

维数是子空间构造的关键，维数过多则次要信息容易引起误分类，维数过少不能体现模式的主要本质特征，也难以将样本正确归类，这里采用最小描述长度(Minimum Description Length，MDL)方法来确定子空间维数[5]

$$\frac{l_{M_j j}}{\sum_{i=1}^{d}l_{ij}}>k_1\geq\frac{l_{M_j+1,j}}{\sum_{i=1}^{d}l_{ij}} \tag{8}$$

式中　$k_1\in(0,1)$是一固定常数，体现了前M个特征值在特征值总和中占的比例。MDL方法的含义是当k_1介于某个特征值及其下一个特征值在d个特征值总和中所占的比例时，该点就是确定特征维数的分界点。

综上所述，改进后的LSM-ALSM算法步骤如下：

1) 计算各类的初始散布矩阵$\boldsymbol{S}^{(j)}(0)$；
2) 求出散布矩阵的特征值和特征向量，按特征值大小排序；
3) 按式(8)确定各类别的子空间维数，由此选择各类别的子空间基矢量；
4) 根据上次误分类的样本，按式(4)旋转各类别的子空间；
5) 根据式(7)和上次分类结果，用性能函数的最佳取值确定拒识规则的参数m_1和m_2；
6) 按式(6)用改进的分类规则，对训练样本进行分类，找出两类错误的样本；
7) 按式(5)调整散布矩阵；
8) 返回步骤2)继续训练，当达到指定的迭代次数或样本的分类精度开始下降为止。

3 识别实验

本文的识别器是针对高校毕业生就业信息光电录入系统所设计，识别对象为26个大写英文字母及数字0~9共36个字符，其中字母O和数字0的识别是借助这两个字符在表格上出现的位置来区分，共收集了每个字符的2 000个样本，1 000个用于训练，1 000个用于测试，测试结果如表1所示。

实验结果显示，改进后的LSM-ALSM算法不仅提高了识别率，而且对可靠性也有所改善。本文提取的字符特征为192维，而LSM-ALSM算法构造出的各类别的子空间为20维左右，即信息量压缩到原来的10%左右，运算速度得到了成倍的提高。

表1 三种子空间分类器的识别结果

识别方法	迭代次数	识别率/(%)	误识率/(%)	拒识率/(%)	可靠性/(%)
LSM	50	87.1	6.2	6.7	93.4
ALSM	100	91.2	4.2	4.6	95.6
LSM-ALSM	50	96.4	1.1	2.5	98.9

4 结 束 语

LSM和ALSM作为两种经典的子空间模式识别方法，各有其优点与不足，本文将LSM的旋转策略引入ALSM的子空间构造中，使LSM的局部调整特性和ALSM的全局平均性有效结合在一起，互为补充，提高了分类器的识别能力。通过对分类函数的加权、子空间维数的合理选择以及性能函数的采用，进一步完善了平均学习子空间分类算法。

参 考 文 献

[1] Oja E. Subspace method of pattern recognition[M]. England: Research Studies Press, 1983

[2] Oja E, Karhunen J. The ALSM algorithm-an improved subspace method of classification[J]. Pattern Recognition, 1983, 16: 421-427

[3] 彭 键. 多类小字符集自适应字符识别技术及系统的研究: [学位论文][D]. 重庆: 重庆大学, 2002, 27-45

[4] 韩 宏, 杨静宇. 神经网络分类器的组合[J]. 计算机研究与发展, 2000, 37(12): 1 487-1 493

[5] Jorma L, Oja E. Subspace dimension selection and averaged learning subspace method in handwritten digit classification[C]. ICANN 1996, 227-232

编 辑 徐培红

·科研成果介绍·

IP电话网关系统

主研人员：李毅超 朱清新 徐 洁 黄克军 张小松 杨小平 宋才栋 李大鹏 胥 能 朱肇乾 赵继东 周金波 文 宇 曾家智 周明天

IP电话网关系统由两个独立的方案成果构成：1) IPS－2000A是一个基于H.323协议的系统，具体实现了全部H.323协议网关功能以及部分Gatekeeper功能和网管功能：内容包括PSTN网和IP网络的H.323协议的信令互译，语音编码解码，业务接续控制，路由选择，网关的配置监视，呼叫记录查询及报表等功能。2) ST2000IP电话网关是根据H.323协议原理改进简化而实现的IP电话网关。利用国产语音卡，提供与电话网和IP网络的互连接口，实现了语音编码解码，通信协议转换，业务接续控制，路由选择，计费数据生成及主叫用户提供语音提示等功能。

·文 争·