基于免疫算法的分类器设计*)

陈云芳　王汝传
（南京邮电大学计算机学院　南京 210003）

摘　要　结合免疫算法强大的自适应识别能力以及全局搜索能力，提出了一种基于人工免疫原理的分类器。根据生物免疫的基因进化、否定选择以及克隆选择原理，建立分类器算法的数学模型并设计了一个基于免疫算法的分类器。最后利用该分类器对手写数字识别问题进行测试。与其他分类算法的实验结果比较表明，该算法在识别率和识别性能方面具备相当竞争力。

关键词　分类器，人工免疫，手写数字识别

Design of Classifier Based on Immune Algorithm

CHEN Yun-fang　WANG Ru-chuan
(Computer Department, Nanjing University of Post and Telecommunication, Nanjing 210003, China)

Abstract　A novel classifier based on immune algorithm was proposed to combine the adaptive recognition capability with global exploration capability. According the principle of gene revolution, negative selection and clonal selection, the mathematical model of classification algorithm was given and a classifier based on immune algorithm was designed. Finally, the algorithm proposed was tested on handwritten digit recognition problem, and the comparison with other classification algorithms indicated that it has a great competitiveness on recognition rate and recognition capability.

Keywords　Classification, Artificial immune system, Handwritten digit recognition

1　引言

分类器在机器学习、模式识别和数据挖掘等领域应用广泛。训练数据集[1]由一组训练样本构成，每个训练样本是一个由特征值组成的特征向量，且每个训练样本还有一个类别标号。给定的训练数据集用来建立一个分类器，所建立的分类器用来判定未知样本的所属类别。一些经典分类器包括：决策树[2]、神经网络[3]、遗传算法[4]、贝叶斯分类[5]、K2 近邻算法[6]、粗糙集[7]和支持向量机[8]。Lim 等人[9]从预测准确度、模型复杂性以及模型训练时间对 33 种分类算法进行了比较。目前普遍认为不存在某种方法能适合各种特点的数据，实际问题的复杂性以及分类方法的本原缺陷都使得无论哪一种方法都只能解决某一类问题。人工免疫系统是一种由生物学启发而来的计算范式，其借鉴了一些免疫学的功能、原理和模型，用于解决复杂问题[10]。许多学者利用生物免疫机制设计出多种免疫算法和模型，其研究成果[11-13]主要涉及机器学习、异常检测、最优化等诸多领域。

生物免疫系统在模式识别方面具有一定的本原特性，比如能够以较少数量级的抗体识别 10^6 倍数量级的抗原[14]，在识别过程中能够保持多样性[15]、抗原的识别效率较高以及具备长期效应能力等等。因此，众多学者开展了生物免疫系统在模式识别中的研究与探索。1993 年，Forrest[16]等人首次采用基于二进制字符串的基因算法研究了生物免疫系统中的两个模式识别问题，讨论了其信息处理与学习机制。2000 年 Dasgupta[17]等人比较了应用于模式识别中的人工免疫否定选择算法和肯定选择算法，并用于检测所监控数据的异常行为。2002 年 de Castro 和 Timmis[18]提出了模式识别中人工免疫计算模型范式，并将其与基于神经网络的模式识别进行深入比较。2005 年 Hart[19]进一步探讨了基于人工免疫的模式识别算法中的识别域和算法参数的选择。目前国内外尚没有深入研究基于免疫系统理论的分类器理论算法与实际应用。

本文提出一种基于人工免疫系统理论的分类器，提取训练数据集统计特征作为受体，采用否定选择算法对训练样本进行有监督训练产生抗体库，再采用克隆选择算法使用抗体库进行未知样本的自适应识别，并在识别过程中不断进化抗体库。本文最后使用该分类器解决手写数字识别问题，验证了算法的准确性和鲁棒性并分析了算法相关参数。

2　系统原理与数学模型

免疫系统的主要工作机制[20]可以归纳为基因库进化、否定选择、克隆选择和免疫记忆 4 个阶段。首先基因库自然进化，从基因片段中生成数量众多、类型不同的候选抗体，这些候选抗体中既包括能够杀死抗原的最终成为真正抗体的细胞，又包括会杀死自身细胞的、对人体有害的细胞。接着进行否定选择，对那些对人体有害的候选抗体进行筛除，只保留那

*)本课题得到国家自然科学基金(60573141 和 60773041)，江苏省高技术研究计划(BG2006001)，国家高科技 863 项目(2006AA01Z201，2006AA01Z219，2006AA01Z439，2007AA01Z404，2007AA01Z478)，2006 江苏省软件专项，南京市高科技项目(2006 软资 105，2007 软资 106，2007 软资 127)，现代通信国家重点实验室基金(9140C1105040805)，江苏省计算机信息处理技术重点实验室基金(kjs06006)和江苏省高校自然科学研究计划(07KJB520083)资助。**陈云芳**　博士生，主要研究方向为计算机软件、计算机网络、信息安全、移动代理等；**王汝传**　教授，博士生导师，主要研究方向为计算机软件、计算机网络和网格、信息安全、无线传感器网络、移动代理和虚拟现实技术等。

些对人体有益的、可能会识别抗原的免疫细胞。然后进行克隆选择，对那些已经成为抗体并且能够有效杀死抗原的细胞进行克隆，以维持正常的抗体数量，保证免疫系统的高效性。最后产生一定的免疫记忆细胞，保证免疫响应能力的长效性。以上过程不断往复并维持某种程度的动态平衡，以保持系统稳定。

人工免疫系统的通用工程框架[21]如图1所示，分别为表示层、亲和力定义和免疫算法层。首先将工程中需要计算的对象以合适的方式表达出来，包括抗原与抗体的定义，接着定义抗原与抗体之间的亲和力，最后选择合适的免疫算法进行相关计算。

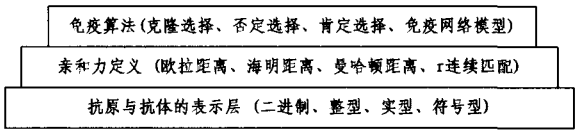

图1 人工免疫系统的工程框架

首先给出分类器的相关定义，假设共有 M 个类别，分别为 $C_1, C_2, \cdots, C_M$。

定义1 抗原结构定义为 n 维向量 $Ag=(ag_1, ag_2, \cdots, ag_n)$，其中 $ag_i \in R, Ag \subset R^n$。

定义2 抗体与抗原的结构相类似，定义为 $Ab=(ab_1, ab_2, \cdots, ab_n)$，抗体和抗原的最佳对应关系是 $Ag=\overline{Ab}$，即 $\forall i \in \{1, \cdots, n\}, ag_i=-ab_i$。

定义3 采用变种的欧拉距离来表示抗原和抗体的结合强度：

$$t=\sqrt{\sum_{j=1}^{n}(Ag_j+Ab_j)^2} \tag{1}$$

其中，Ag_j 表示抗原向量的第 j 个分量，Ab_j 表示某个抗体向量的第 j 个分量，n 为向量维数。

定义4 亲和力 A 表达抗原和抗体之间的相似性：

$$A=1/(1+t) \tag{2}$$

A 的值在0和1.0之间。A 值越大，表示抗原和抗体越匹配。$A=1.0$ 时，表示抗体与抗原理想结合，得到最优匹配。

定义5 抗体数目上限 MaxNum。为了控制每个类别所对应的抗体数目，设置了所有抗体数目的总上限，用于调节识别率和执行性能之间的平衡：

$$\text{MaxNum}=N_{C_1}+N_{C_1}+\cdots+N_{C_M} \tag{3}$$

其中 N_{C_i} 标识类别 C_i 的抗体数量。

定义6 抗体耐受阈值 Tolerance_t。初始抗体库生成过程中以及抗体库进化过程中，新产生抗体 Ab' 与抗体库中的原有抗体 Ab 的亲和力大于免疫耐受阈值时，则产生免疫耐受而不接受该抗体加入原抗体库。

$$\forall C_i \in \{C_1, C_2, \cdots, C_M\}, \forall i \in \{1, \cdots, N_{C_i}\}$$
$$\frac{1}{1+\sqrt{\sum_{k=1}^{M}Ab_{ik}-Ab_{jk}'}} > Tolerance_t \tag{4}$$

定义7 识别阈值 Recognize_t。当被识别未知样本的特征向量所对应的抗原与抗体库中的抗体经过计算后，其亲和力 A_i 都小于最小识别阈值时，则拒绝识别该未知样本。即若 $\forall i \in \{1, \cdots, \text{MaxNum}\}, A_i < Recognize_t$，则拒绝识别该样本。

综合以上定义，若对于未知样本 $K, A_i > A_j > Recognize_t, i, j \in \{1, \cdots, \text{MaxNum}\}$，则该样本 K 属于抗体 i 所对应的那个类别。整个训练和识别过程可描述如下：首先采用训练样本集进行有监督学习训练，产生初始抗体库，该抗体库中每个类别都有与其对应的、数量不等的抗体，这些抗体之间互相免疫耐受，且其总数目不超过抗体数目上限。进行未知样本识别时，先抽取样本基本特征，产生对应的抗原，将此抗原与抗体库中的抗体进行逐条亲和力计算。在不低于最低识别阈值的前提下，取亲和力最大的类别作为识别结果。

3 分类器算法描述

与生物免疫过程的否定选择产生抗体和克隆选择抗体进化过程类似，本系统的算法主要由两个部分组成：其一是抗体库的产生算法，其二是免疫识别算法。

算法1 抗体库的产生算法

输入：训练样本集； 输出：初始抗体库

①抗体库文件记录清空，打开训练样本集

②读取训练样本集中一个样本，依据样本特征向量产生对应的抗体

③计算该抗体与抗体库中该对应类别已有抗体的耐受力

④若免疫耐受力超出，则忽略该抗体，若免疫耐受力低于阈值，则：

a. 若抗体总数目小于抗体数目上限，则直接将该抗体加入至抗体库中

b. 若抗体总数目等于抗体数目上限，则替换其中最近最少被使用的抗体

⑤重复步骤②－④，直至训练样本集为空

算法2 免疫识别算法

输入：未知样本； 输出：识别结果，进化的抗体库

①打开抗体库文件，产生抗体数据链表

②特征化未知样本，产生未知样本对应的抗原

③取一条抗体记录，计算抗原与该抗体的亲和力

④重复步骤③，记录下亲和力最大值以及其对应的抗体类别

⑤如果该亲和力最大值小于识别阈值，那么就拒绝识别该未知样本，否则将该样本识别为抗体对应类别

⑥如果出现识别错误，则进行有监督的学习，根据教师判定类别将正确识别结果所对应抗原参与抗体库的进化，过程类似于抗体库的初始生成

算法的时间复杂度主要取决于两个因素，抗体数目上限和抗体向量维数。另外，样本类别数和抗体表达方式也影响算法的计算量。

4 实验与结果分析

我们采用模式识别中的一个经典手写数字识别问题作为本算法的测试用例。相对于其他模式识别问题而言，手写数字识别的模式特征具有一定的聚集性，其分类器所采用的方法非常多[23]，比如人工神经网络（ANN）[24,25]、支持向量机（SVM）[26]、隐马尔可夫模型（HMM）[27,28]、主成分（PCA）[29]、主曲线（PC）[30]等方法以及几种方法的集成[31,32]。

我们采用美国邮政局采集的 USPS[33] 手写数字集数据库进行测试，其中训练集包含7291个数字的特征，测试集包含2007个数字的特征。实验首先使用训练集作为基因片段进行抗体库的训练，采用否定选择算法产生初始抗体库。接着，使用克隆选择算法对能够识别手写数字的抗体进行简单克隆，以维持正常抗体活性。每次手写数字的识别过程既是产生识别结果的过程，也是抗体库不断进化的过程。不同用户使用该系统后会导致抗体库产生一些差异，既能识别大多数人所书写的数字，也能识别最近经常使用系统的用户所书写

的数字。

根据手写数字识别的特点，我们对本文算法中进行如下设置：①采用一个实数向量集合表示抗体和抗原，以提高区分度；②降低交叉反应阈值以提高识别精度，同时增加抗体数目以维持较高的识别率；③采用复杂度较低的亲和力匹配算法，以提高识别速度；④降低基因库进化频度，因为相比较而言，手写数字模式差异有限，抗体库无需频繁更新。

将字符分割为 $m*m$ 的小块，计算每个小块内字符像素的个数，除以小块的面积，得到字符在该块位置的特征 ag_i。生成初始抗体库的算法所需要的时间与所采用的训练集大小有关，若训练集大小为 10^3 数量级，则所需时间约为数分钟。手写数字识别算法的时间复杂度为 $O(MaxNum*m^2)$，实际测试时无明显延时。将抗体库大小设定为 300 条，每条记录存放 $m*m$ 个浮点数。若每个浮点数占 4 个字节，m 为 16，则抗体库文件最大约为 300k 字节。

分类器性能最一般使用正确识别率来度量，$C_r=n_r/N$，其中 n_r 为正确分类的样本数，N 为样本总数。为了得到更好的识别率和识别性能，我们通过调整字符切分维数、抗体耐受阈值、抗体数目上限和拒识阈值等系统参数比较识别结果，来进一步分析算法的性能。

4.1 切分维数对识别率的影响

m 是系统的主要参数之一，m 越小，表明切分的块数越少，取得的字符特征就越少，抗原之间的区分度就越小。m 越大，取得的字符特征就越多，抗原之间的区分度就越大，但系统计算的负担就越重。图 2 可以看出系统的识别率总体趋势是随着 m 的增大而提高。实验同时还表明对简单特征提取的手写字符识别而言，切分维数为 16 是足够充分的。

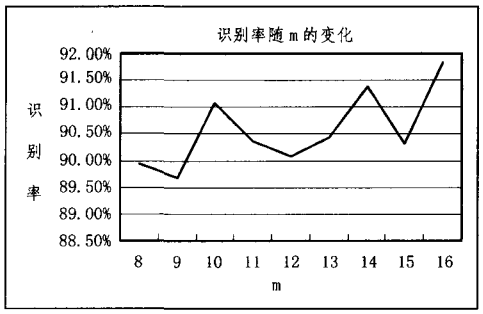

图 2 识别率与切分维数之间的关系

4.2 抗体数目上限对识别率的影响

当切分维数为 16，耐受阈值为 0.9，识别阈值取 0.75 时，由表 1 看出，增大抗体数目上限 MaxNum，系统识别率将随之提高。但系统识别率增长速度随着 MaxNum 增大而趋缓，MaxNum=250 和 MaxNum=300 具有相同的识别率。虽然系统可以储存的抗体数目增多了，但由于抗体的耐受功能和某些数字写法变化不多，训练后这些数字的抗体数目依旧较少。

表 1 抗体数目上限 MaxNum 对识别率的影响

MaxNum	识别率(%)
50	88.99
150	92.63
250	93.27
300	93.27

4.3 识别阈值对识别率的影响

识别阈值也是系统非常重要的一个参数。如表 2 所示，识别阈值太低，则误识率会比较高；识别阈值太高，则拒识率比较高，实验表明，当切分维数为 16，耐受阈值为 0.9，抗体数目上限为 300 时，识别阈值取 0.75 比较适中。

表 2 识别阈值对识别率的影响

识别阈值	识别率(%)
0.65	89.54
0.75	93.27
0.85	91.28

4.4 训练样本数目对识别率的影响

当切分维数为 16，耐受阈值为 0.9，抗体数目上限为 300 时，识别阈值为 0.75。我们分别采用不同数量的训练样本来产生抗体库。实验表明，我们的算法能够在有监督的小样本训练下达到比较理想的识别率(表 3)。

表 3 训练样本数对识别率的影响

训练样本数	测试集数目	正确识别数目	识别率(%)
1000	2007	1634	81.42
2000	2007	1828	91.08
3000	2007	1870	93.17
7291	2007	1872	93.27

4.5 与其它识别算法的比较

最后，我们设定如下参数：抗体上限数目为 300，免疫耐受阈值为 0.9，拒识阈值为 0.75，切分维数为 16×16。针对 USPS 数据集，分别采用不同的识别算法来比较它们的识别率。

传统 BP 神经网络(BPN)采用 3 层前向神经网络，输入层节点为 9，隐含层节点为 19，输出层节点为 10，采用加入动量项的学习速率渐小算法，其中动量项 $A=0.7$，初始学习率 $G_0=0.01$，网络收敛误差 $E=0.001$。

基于最佳鉴别方向的隐马尔可夫过程(HMM)[34]采用 9 个最佳鉴别方向上编码，利用 Baum-Welch 算法建立转移状态数为 6 的 HMM 模型，最后通过 Forward-Backward 算法计算测试样本条件概率，实现数字识别。

基于主成分分析(PCA)[35]采用 PCA 对字符特征进行特征抽取和降维，选取其中的 6 个主成分，使得抽取的主成分具有较稳定的模式特征和较低的维数，并以此估计字符重建模型，最后通过重建误差的分析实现字符识别。

表 4 不同识别方法的识别率比较

	测试集数目	正确识别数目	识别率(%)
BPN	2007	1690	84.21
HMM	2007	1850	92.17
AIS	2007	1872	93.27
PCA	2007	1898	94.57

实验表明，AIS 算法具有很强的竞争性，能够以较小的计算时间复杂度和空间复杂度达到比较理想的识别性能。其中基于主成分分析(PCA)的识别算法虽然识别率高于本文描述的算法，但是其 6 个主成分的计算量已经远远超出了本算法的时间复杂度。

相对于其它的识别算法[24-28]，本算法具有生物免疫系统所带来的优良特性。

①由于系统具备一定的进化能力，学习样本的规模可以较小。系统能在不断使用的过程中不断进化抗体库，以维持较高的自适应识别能力。

②由于每个数字均存在多个互相免疫耐受的抗体，抗体

库中的抗体的总数量可以在较高识别率的前提下维持在比较少的水平，并能够避免出现过拟合现象。

③系统不需要复杂的手写数字特征提取方式和较复杂的特征匹配算法即可实现较高的识别率，识别性能相对较高。

④系统具备较高的并行度，各个类别的亲和力计算过程相对独立，很容易实现并行计算，从而进一步提高整体性能。

结束语 本文提出了一种基于人工免疫思想的分类器算法，描述了该算法数学模型的相关定义及演算步骤，并根据该算法实现了一个手写数字识别系统，证明了基于人工免疫原理分类器的理论可行性，最后使用 USPS 数据进行了实际测试，实验表明该算法在分类器算法中具有很强的竞争优势。

本算法还存在一些不足和改进之处，亦是未来深入研究之方向[36]：样本特征的提取在整个识别过程中起重要作用，免疫系统优秀特性的发挥还依赖于抗原表达的精确细致化。改进抗体库的产生和进化过程，进一步在不牺牲执行性能的前提下提高抗体的有效覆盖率，比如识别阈值为可变长。根据抗原表达定义适合的亲和力计算方法。每个类别的信息熵值通常不一致，根据识别率的高低动态调整各个类别的抗体数目。

参 考 文 献

[1] 李宏东，姚天翔. 模式分类. 第二版. 北京：机械工业出版社，2003

[2] Quinlan J R. Induction of decision trees. Machine Learning，1986 (1)：81-106

[3] Rumelhart D E，Hinton G E，Williams R J. Learning internal representations by error propagation. Parallel Distributed Processing. Cambridge，MA：MIT Press，1986

[4] DeJong K A，Spears W M，Gordon D F. Using genetic algorithms for concept learning. Machine Learning，1993，13：161-188

[5] Langley P，Iba W，Thompson K. An analysis of Bayesian classifiers// AAAI(1990). 1990：223-228

[6] Cover T M，Hart P E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory，1967，13：21-27

[7] Pawlak Z. Rough Classification. Int. J Man Machine Studies，1984，20：469-483

[8] Cristianini N，Shawe T J. An Introduction to Support Vector Machines. Cambridge University Press，2000

[9] Lim T，Loh W，Shih Y. A Comparison of Prediction Accuracy，Complexity，and Training Time of Thirty Three Old and New Classification Algorithms. Machine Learning，2000，40：203-228

[10] de Castro L N，Timmis J I. Artificial Immune Systems：A New Computational Intelligence Approach. Springer-Verlag，2002

[11] Dasgupta D，et al. Artificial Immune System (AIS) Research in the Last Five Years // IEEE Conference on Electronic Commerce. 2003

[12] Hart E，Timmis J. Application Areas of AIS：The Past，The Present and The Future// International Conferences on Artificial Immune Systems 2005. Springer-Verlag，2005

[13] de Castro L N，Timmis J I. Artificial immune systems as a novel soft computing paradigm. Soft Computing Journal，2003，7(8)：526-544

[14] Parkin J，Cohen B. An overview of the immune system. The Lancet，2001，357：17

[15] Oprea M，Forrest S. How the immune system generates diversity：Pathogen space coverage with random and evolved antibody libraries// 1999 Genetic and Evolutionary Computation Conference (GECCO). July 1999

[16] Forrest S，Javornik B，Smith R，et al. Using genetic algorithms to explore pattern recognition in the immune system. Evolutionary Computation，1993，1(3)：191-211

[17] Dasgupta D，Nino F. A comparison of negative and positive selection algorithms in novel pattern detection// The Proceedings of the IEEE International Conference on Systems，Man and Cybernetics (SMC). Nashville 2000

[18] de Castro L N，Timmis J. Artificial immune systems：a novel paradigm for pattern recognitionArtificial Neural Networks in Pattern Recognition. 2002：67-84

[19] Hart E. Not all balls are round：An investigation of alternative recognition-region shapes// International Conferences on Artificial Immune Systems. 2005：29-42

[20] Timmis J，Knight J I，de Castro T L N，et al. An Overview of Artificial Immune Systems：An Emerging Technology. invited chapter for the book CYTOCOM，2001

[21] Dasgupta D. Immunity - based Intrusion Detection Systems：A General Framework// the Proceedings of the 22nd National Information Systems Security Conference (NISSC). October 1999

[22] Perelson A S，Weisbuch G. Immunology for physicists. Reviews of Modern Physics，1997，69：1219

[23] Dong J X，Krzyzak A，Suen C Y. Comparison of algorithms for handwritten numeral recognition. Technical Report，CENPARMI. Montreal：Concordia University，1999

[24] LeCun Y，Boser B，Denker J S，et al. Handwritten digit recognition with a back-propagation network. Advances in Neural Information Processing Systems，1990(2)：396-404

[25] Lee S-W. Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network. IEEE Trans. Pattern Anal. Machine Intell.，1996，18：648-652

[26] Bahlmann C，Haasdonk B，Burkhardt H. Online handwriting recognition with support vector machines-a kernel approach // Frontiers in Handwriting Recognition，2002. Proceedings. 8th International Workshop. 2002：49-54

[27] HuJ，Brown M K，Turin W. HMM based on - line handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intel.，1996，18：1039-1045

[28] Schlapbach A，Bunke H. Off-line Handwriting Identification Using HMM Based Recognizers // 17th International Conference on (ICPR'04). Pattern Recognition，Cambridge UK，2004：654-658

[29] Deepu V，Madhvanath S，Ramakrishnan A G. Principal Component Analysis for online handwritten character recognition // ICPR 2004，Proceedings of the 17th International Conference on Pattern Recognition. Aug. 2004，2：23-26

[30] Kegl B，Krzyzak A，et al. Piecewise linear skeletonization using principal curves. IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，24 (1)：59-74

[31] 苗夺谦，张红云，李道国，等. 基于主曲线的脱机手写数字识别. 电子学报，2005，33(9)：1639-1643

[32] Salah A A，Alpaydin E，Akarun L. A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，24 (3)：789-796

(下转第 174 页)

表 1 “启动电子计时器”中“qi”的识别过程

	1	2	3	4	5
HMM 概率最大的前 5 个音	zi	qi	ying	ji	ni
HMM 下的条件概率	0.393	0.353	0.225	0.026	0.01
拼音出现在句首的概率	0.01	0.025	0.015	0.035	0.045
HLM 概率	0.0039	0.0088	0.0033	0.0009	0.0004

表 2 “dong”的识别过程

	1	2	3	4	5
HMM 概率最大的前 5 个音	duo	dong	zuo	chong	huo
HMM 下的条件概率	0.540	0.272	0.116	0.070	0.01
拼音在“qi”之后的概率	0.01	0.047	0.01	0.01	0.01
HLM 概率	0.0054	0.012	0.0011	0.0007	0.0001

显然，利用语言模型的先验知识后，一些 HMM 下概率最大，但是错误的候选音得到了一定的校正。

4.2 语言模型背景下的文字流解析

图 5 给出了 HLM 方法的音识别率和在此基础上利用语言模型进行文字流解析的实验结果。数据库中有 935 个汉字、96 个音，如不借助语言的先验知识，平均每个音对应 10 个汉字，盲目解析的正确率只有 10%。而应用语言模型的正确率最高可达 70.5%，对单一视觉流的文字解析而言是可观的。这充分说明了语言模型在文字流解析环节的强大作用。

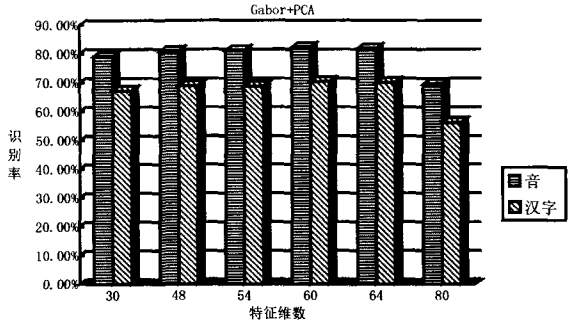

图 5 音识别与文字流解析性能比较

综上，语言模型对唇读的拼音识别和文字流解析环节都有积极的作用。但是，简单的语言模型存在一些问题，即当某个音识别错误时，很可能紧邻其后的音也被识别错误，类似于“多米诺骨牌效应”，尤其是句首音误识时，表 3 是一个例子。因此，语言模型有待我们改进成可回溯的模型，比如，加入反馈校正环节。

表 3 “类多米诺骨牌效应”

正确汉字	正确拼音	HMM	HLM
吃	chi	chi	shi
点	dian	dian	ye
东	dong	nong	nong
西	xi	xin	xin

结束语 本文针对唇读中口型序列和语言序列的一对多映射问题，主要研究语言模型对唇读的作用，突破单纯采用声学后验概率进行识别的传统框架，建立融合 HMM 和语言背景知识的新模型 HLM，并应用语言模型进行文字流解析。实验表明，建立语言模型能校正部分 HMM 识别错误的音，对单一视觉流的说话内容识别和文字流解析起着重要积极作用。它是由唇动序列得到语言序列唯一解的有效途径，必然成为未来实用唇读系统中不可或缺的一部分。语言模型在唇读中的应用虽有不少问题尚待解决，但并非不可解决。比如，解决“类多米诺骨牌效应”问题，可通过加入回馈环节等。

参 考 文 献

[1] Potamianos G, et al. Audio-Visual Automatic Speech Recognition: An Overview [M]. MIT Press, 2004

[2] Potamianos G, Graf H P, Cosatto E. An Image Transform Approach for HMM Based Automatic Lipreading[C]// Proc. Int. Conf. Image Processing. 1998, 1: 173-177

[3] Potamianos G, Neti C. Improved ROI and Within Frame Discriminant Features for Lipreading[C]// Proc. Int. Conf. Image Processing. Thessaloniki, Greece, 2001, 3: 250-253

[4] 姚鸿勋，高文，王瑞，等. 视觉语言——唇读综述[J]. 电子学报，2001, 29(2): 239-246

[5] Potamianos G, et al. Recent Advances in the Automatic Recognition of Audio-visual Speech[C]. Proc. of the IEEE, 2003, 91(9): 1306-1326

[6] Rosenfeld R. A Maximum Entropy to Adaptive Statistical Language Learning[C]. Computer Speech and Language, 1996, 10(3): 187-228

[7] Chomsky N. Aspects of the Theory of Syntax [M]. Cambridge: MIT Press, 1965

[8] Chomsky N. Syntactic structures [M]. Mouton, 1964

[9] 黄昌宁，张小凤. 自然语言处理技术的三个里程碑[J]. 外语教学与研究，2002, 34(3): 180-187

[10] 王晓龙，关毅. 计算机自然语言处理[M]. 北京：清华大学出版社，2005: 47-68

[11] Hong Xiaopeng, Yao Hongxun, Wan Yuqi, et al. A PCA Based Visual DCT Feature Extraction Method for Lip-reading[C]// Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing. 2006

[12] Katz S M. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer[C]. IEEE Transaction on Acoustic, Speech and Signal Processing, 1987, 35(3): 400-401

[13] Kneser R, Ney H. Improved Backing-off for M-gram Language Modeling// Proceedings of the IEEE[C]. Int. Conf. on Acoustics, Speech and Signal Processing. Detroit, MI, USA, 1995: 181-184

[14] Rosenfeld R. Two Decades of Statistical Language Modeling: Where Do We Go from Here? [C]// Proceedings of the IEEE. 2000, 88(8): 1270-1278

(上接第 136 页)

[33] Dong J X, Krzyzak A, Suen C Y. Statistical result of human performance on USPS database. Technical Report, CENPARMI. Concordia University, 2001

[34] 芮挺，沈春林，丁健. 基于最佳鉴别变换的 HMM 手写数字字符识别. 中国图像图形学报，2004, 9(8): 1008-1013

[35] 芮挺，沈春林，丁健，等. 基于主分量分析的手写数字字符识别. 小型微型计算机系统，2005, 26(2): 289-292

[36] Garrett S M. How Do We Evaluate Artificial Immune Systems? Evolutionary Computation, 2005, 13(2): 145-178