手语识别中基于 HMM 的区分性训练方法

王雨轩, 倪训博, 姜 峰
(哈尔滨工业大学 计算机学院, 黑龙江 哈尔滨 150001)

摘 要:传统的隐马尔科夫模型(HMM)的训练方法基于统计概率的最大似然准则(MLE),在训练样本数目足够大的情况下,这种方法在理论上可以得到最优的结果. 在手语识别研究中,采集足够大的训练样本十分困难. 区分性训练可以很好地弥补由于训练样本的缺乏以及手语模型之间的近似而造成的识别系统的缺陷. 最大交互信息准则(MMIE)作为区分性训练准则的一种已经被广泛的应用于语音识别领域. 文中通过合理的构建手语识别中的竞争模型和易混集,提出了 MMIE 准则的改进形式,并将其应用于特定人与非特定人手语识别. 实验证明,使用改进的 MMIE 准则对识别系统性能有很大的提高.

关键词:区分性训练;隐马尔科夫模型;易混集;最大交互信息
中图分类号: **文献标识码**:A **文章编号**:1673-4785(2007)01-0080-05

Discriminative training methods of HMM for sign language recognition

WANG Yu-xuan, NI Xun-bo, JIANG Feng
(School of Computer Science, Harbin Institute of Technology, Harbin 150001, China)

Abstract: The traditional method of training HMM (Hidden Markov Models) is based on MLE (maximum likelihood estimation). When training samples are sufficient enough, the method can principally gain the optimal result. However, it is too difficult to get such large data sets practically, especially in sign language recognition. Discriminative training method can improve the error rate of MLE, which is caused by insufficient training data and similarities among sign language models. Maximum mutual information estimation as one of discriminative training methods has been widely applied in speech recognition. By taking competition models into account and setting up mixture sets appropriately, MMIE method was improved and applied both in signer-dependent and signer-independent sign language recognition. A great number of experiments had been taken, showing that this method greatly promoted the ability of the traditional MLE system.

Keywords: discriminative training; hidden Markov models; mixture sets; maximum mutual information

手语作为一种结构化手势,是聋人进行信息交流的最常用形式. 自动手语识别的尝试始于 20 世纪 90 年代. 新加坡南洋理工大学 Charayaphan 和 Marble[1] 使用图像处理方法来理解美国手语中 31 个孤立手势词,该方法能正确识别其中的 27 个. 此后,国际上众多学者投入到手语识别的领域中,比较著名的如香港中文大学 Deng 和 Tsui[2] 使用基于并行的 HMM 模型去识别 192 个美国手语词,识别率为 93.3 %.

目前手语识别研究中,最常用的是基于高斯混合概率密度的 HMM 模型系统:采用传统的 MLE 准则函数与 BW(Baum-Welch)算法对模型的各个参数进行迭代重估. 这种重估方式只考虑当前模型的所有训练样本,不考虑模型之间的相关性.

最大交互信息准则 MMIE[3],是最为常用的区分性训练准则. 与 MLE 相比,MMIE 在训练时不仅考虑到当前模型的信息,还考虑到其他竞争模型的信息. 这就使 MMIE 准则可以很好地作为 MLE 训练准则的补充.

在语音识别领域,对区分性训练的方法进行了

收稿日期:2006-04-29.

广泛的研究,Normandin[4]等人采用 EBW(expend-baum-welch)算法实现了 MMIE 准则在连续 HMM 模型中训练的难题,使这种方法开始广泛应用于语音识别领域.

在手语识别领域,由于易混集的构建等问题,区分性训练还没有被手语识别研究者所采用.然而由于手语信号和语音信号都是基于统计概率的时序信号,可以期待对于 MMIE 准则在手语识别领域上的改进能够极大地改善现有系统的识别效果.

1　MMIE 准则及对其的改造

1.1　MMIE 准则的基本原理

传统的 MMIE 的目标函数为

$$F_\lambda = \sum_r^R \ln \frac{P(O^r \mid \lambda)}{\sum_m^{M_r} P(O^r \mid \lambda_m)}. \quad (1)$$

式中:R 为当前训练的样本个数,M_r 为由当前词产生的易混词表,λ 为正确 HMM 模型所对应的参数,λ_m 为易混词表中的一个 HMM 模型所对应的参数.

而传统的 MLE 目标函数为

$$F_\lambda = \sum_k^K \ln P(O^k \mid \lambda). \quad (2)$$

通过对 2 种准则目标函数的比较,MMIE 准则只比 MLE 准则多了分母上的一项易混集上的后验概率的累加.这反映 MMIE 目标函数的本质是增加当前模型的后验概率在易混集中所占的比例,使相近的模型之间的距离增大,以此增强模型的泛化能力,提高识别效果.相对的,MLE 准则只关注于当前训练模型上的所有训练样本的极大似然概率值而忽视了其他近似模型的训练.这就是 2 种训练准则本质上的不同.MLE 在 400 词集上进行训练和在 4 000词集上进行训练得到的训练模型结果都是相同的,因为 MLE 准则下,模型的训练是独立的,非相关的.而 MMIE 准则在 MLE 的基础上考虑模型之间的相关性,这就注定了 MMIE 准则可以很好地弥补 MLE 准则的固有缺陷.二者合作使用,理论上会使结果更加优化.

传统的 MMIE 准则的训练方法是扩展的 BW 算法 EBW.其对 HMM 模型均值和方差的重估公式如下:

$$\mu_g = \frac{\{\theta_g^{num}(O) - \theta_g^{den}(O)\} + D\mu_g}{\{\gamma_g^{num} - \gamma_g^{den}\} + D}. \quad (3)$$

$$\sigma_g^2 = \frac{\{\theta_g^{num}(O^2) - \theta_g^{den}(O^2)\} + D\{\mu_g^2 + \sigma_g^2\}}{\{\gamma_g^{num} - \gamma_g^{den}\} + D} - \mu_g^2. \quad (4)$$

式中:上标 num 和 den 分别对应于当前的模型和所有易混集中的模型.常数 D 用来保证参数计算的结果为正值,同时控制收敛速度.

更一般形式的 MMIE 准则即 H 准则目标函数定义如下:

$$F_\lambda = \sum_r^R \ln \frac{P(O^r \mid \lambda)}{\left(\sum_{m=1}^{Mr} P(O^r \mid \lambda_m)\right)^h}. \quad (5)$$

H 准则目标函数比传统的 MMIE 准则增加了分母上的指数项,这就使 H 准则更具有了一般性.可以看出,MLE 和 MMIE 准则都可以理解为特殊情况下的 H 准则,即当 $h=0$ 和 $h=1$ 时.[0,1]被普遍的认为是 h 的正常、合理的值域,在实验中却发现为了寻求最小错误率,值域的范围可以扩展到 $[1,+\infty)$.

对于 H 准则重估函数的推导可以用传统的 EBW 算法进行扩展或使用改良的梯度下降算法(GD)进行推导.二者推出的重估公式极为相近,如下所示:

$$\mu_g = \frac{\{\theta_g^{num}(O) - h\theta_g^{den}(O)\} + D\mu_g}{\{\gamma_g^{num} - h\gamma_g^{den}\} + D}. \quad (6)$$

$$\sigma_g^2 = \frac{\{\theta_g^{num}(\sigma^2) - h\theta_g^{den}(\sigma^2)\} + D\sigma_g^2}{\{\gamma_g^{num} - h\gamma_g^{den}\} + D}. \quad (7)$$

可见,除了引入 h 系数以及方差重估公式的略微不同外,H 准则重估公式和标准 MMIE 重估公式十分相近.然而适当的设定 h 值可以提高 MMIE 准则的收敛速度及获得更好的识别结果.

1.2　H 准则重估公式在手语识别应用中的改进

H 准则重估公式和 MMIE 准则面临一个同样的问题,即 D 值的选取.通常采取以下 2 种策略:1)使用一个全局最大化的阈值;或者选取以下二者的最大值[5]:1) $h\gamma_g^{den}$;2)使高斯模型各维变量均为正值的最小 D 值的 2 倍.

使用全局最大化阈值相对简单,然而过大的阈值对于易混模型会造成收敛速度下降,以至效果不明显.而使用局部计算 D 值的方法,会造成计算量增加,尤其是计算方差时,通常都要解二次方程,这使算法的复杂度进一步上升.而且由于 D 值选取的不均衡,会对不同的模型造成不良影响.

通过将 H 准则应用于手语识别的大量实验,发现在 h 值选择恰当时,由于易混集构建的特点,完全可以取消常数 D,就可以保证绝大多数参数的结果为正值.由于易混集选中的模型打分基本都高于待训练的模型,所以可以期待分子分母中的易混模型集合的累加项要大于该模型的那一项,分子分母基本上同时为负值,结果为正值.对于少数结果为负值的重估结果,只要简单的将其取反就可以保证其对

结果不会产生很大影响.这样做的好处,一是大大降低了由于求 D 值所造成的复杂度;二是在识别效果上看,由于避免了 D 值选择不当所造成的影响,效果比传统的 EBW 推导出的算法要优越很多.

H 准则重估公式与手语识别中采用的混合高斯 HMM 相结合并采取上述的改造后,均值和方差重估公式结果如下:

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{Tr}\gamma_t^{(r)}(j,m)\cdot O_t^{(r)} - h\sum_{r=1}^{R}\sum_{u\in M}\sum_{t=1}^{Tr}\gamma_{t,u}^{(r)}(j,m)\cdot O_t^{(r)}}{\sum_{r=1}^{R}\sum_{t=1}^{Tr}\gamma_t^{(r)}(j,m) - h\sum_{r=1}^{R}\sum_{u\in M}\sum_{t=1}^{Tr}\gamma_{t,u}^{(r)}(j,m)}. \tag{8}$$

$$\sigma_{jm}^2 = \frac{\sum_{r=1}^{R}\sum_{t=1}^{Tr}\gamma_t^{(r)}(j,m)\cdot\delta_{jmt}^{(r)} - h\sum_{r=1}^{R}\sum_{u\in M}\sum_{t=1}^{Tr}\gamma_{t,u}^{(r)}(j,m)\cdot\delta_{jmt}^{(r)}}{\sum_{r=1}^{R}\sum_{t=1}^{Tr}\gamma_t^{(r)}(j,m) - h\sum_{r=1}^{R}\sum_{u\in M}\sum_{t=1}^{Tr}\gamma_{t,u}^{(r)}(j,m)}. \tag{9}$$

式中:$\delta_{jmt}^{(r)} = (O_t^{(r)} - \mu_{jm})(O_t^{(r)} - \mu_{jm})'$,$R$ 为训练样本个数,Tr 为第 r 个训练样本的帧数,定义第 r 组训练样本的第 t 帧观测到的数据来自状态 S_j 的第 m 个混合分量模型的条件概率密度 $\gamma_t^{(r)}(j,m)$. M 为易混集,$\gamma_{t,u}^{(r)}(j,m)$ 表示第 u 个竞争模型对应的 $\gamma_t^{(r)}(j,m)$. $O_t^{(r)}$ 表示第 r 组样本第 t 帧的观测数据值. h 为 H 准则中的系数.

实验证明,对于 HMM 模型的其他参数,如混合比,转移概率等,参加区分性训练,对结果影响不大.Jing Zheng[6] 等前人的工作证明了均值与方差在区分性训练中起决定性的作用.为了降低时间付出,文中采用均值和方差作为区分性训练的目标参数.

2 易混集的构造及应用

通过对 MMIE 准则的研究可以得到,MMIE 准则是不可以孤立运行的,它需要拥有 ML E 准则所不具备的一些额外信息 —模型之间的交互.这些交互信息可以通过应用已构建好的 ML E 系统产生易混集的形式来实现.反过来这些交互信息进行区分性训练后,就可以提高 ML E 系统的性能[7].同时 MMIE 训练模型的出发点也应该是 ML E 已经构造好的 HMM 模型.

易混集中的竞争模型是和当前模型 ML E 打分相近的模型.这些易混淆的模型有可能是训练数据的缺乏造成,也可能是由于模型本身固有的相似性造成.如何挑选入选易混集的正确模型,以及其相应的竞争模型,对于区分性训练来说是至关重要的一环.

传统构建易混集的方法有 N-BEST 方法等[8-10],这些方法在已有的 ML E 模型基础上,需要对模型进行全局区分性训练,并且机械的选择 N 个打分最高的模型构建易混集,缺乏灵活性,造成训练算法的高复杂度.姜峰等人在基于支持向量机的二层 HMM 模型上对于易混集的构造进行了相应的研究,对文中易混集构造有一定借鉴意义.

对于手语识别中的数据进行大量测试后发现,不同的手语者之间,有一类词经常被误识,比如"百"这个词经常被误识为"八十","百合"等词,构造易混集的目的就是要找出这些词,使区分性训练有的放矢.

构造和使用易混集表的算法:

1) 把原来统一的 ML E 训练数据集按不同手语者进行分组,然后以某个手语者作为测试集,其他剩余的数据作为新的训练集进行 ML E 训练.

2) 以新生成的测试集对新的 HMM 模型进行测试,给出 ML E 打分,将被误识的词记录下来,并将比该词打分高的所有词所对应的模型列入该词的易混集中,作为该词所对应的 HMM 模型的竞争模型.

3) 对于不同手语者均作以上操作,得到一系列的易混集表.

4) 对不同的易混集进行合并:对两两易混集中的被记录的误识词取交集来体现误识模型的共性,对该误识词所对应的竞争模型取并集以综合改误识词针对不同手语者的个性.

5) 以合并后易混集作为交互信息,以原统一的 ML E 训练数据集构造的 HMM 模型为出发点,以上文得到的区分性训练方法的重估公式重新计算 HMM 模型的均值和方差,得到新的 HMM 模型.

6) 用原测试集对新的 HMM 模型进行测试.

与其他的易混集构造策略相比,这种构造易混集的方法非常灵活.构造的易混集完全来自于原训练集,并没有加入测试集的任何信息.由于只考虑误识词,因此大大降低了运算的时间,而实验表明对识别率没有很大的影响.对于不同易混集进行有选择性的合并,既控制了易混集的规模,又选择了相对有效的交互信息.

此外,在对易混集的合并过程中加入一些主观经验知识,将会对识别结果产生积极的影响.

3　实验结果及分析

3.1　实验架构

文中使用具有代表性的 400 手语孤立词汇集,数据由 6 位专业手语老师通过数据手套采集得来.每位手语老师采集 2 遍,一共是 12 遍数据.每词按词本身的结构,打手语者的习惯,以及打手语时的环境不同而有不同的帧长,每一帧有 51 维的观察值.

对于注册集测试,采用手语者的一遍数据加入训练集,另一遍数据作为测试集的方法.对于非注册集测试,则将所有其他人的数据作为训练集,该手语者的一遍数据作为测试集的方法.

将首先针对新的重估公式进行测试,以获取最佳参数,迭代次数等的信息.最后给出在不同易混集构建方法下,MLE,EBW 算法,改良的算法的识别率的比较.

3.2　h 参数及迭代次数的分析

在 H 准则中,h 参数对于重估公式影响很大.它既关系到收敛的速度,又关系到识别率.经过大量的实验,发现将 h 值从传统的 $[0,1]$ 区间扩展到 $[1,+\infty)$ 会获得更好的效果.以注册集上的区分性训练一次迭代后的结果为例,如图 1,在 h 值定为 1.7 时,效果最好.在其他训练情况下,也有类似结果.

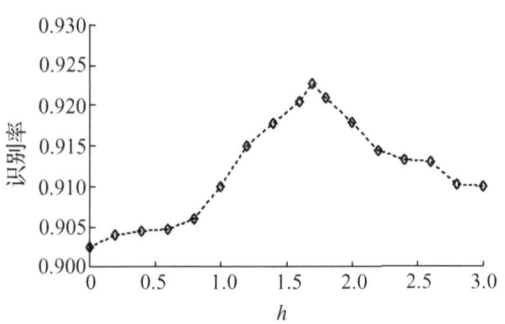


图 1　参数 h 的实验

Fig. 1　The experiment of h

对于区分性训练的迭代次数也进行了大量的实验,结果发现在第 4 次或第 5 次迭代时,会得到很好的收敛效果,再继续训练,将会造成发散.

如图 2 可以看到 MIE 算法已经收敛到极限,对其进一步的迭代计算将不会对结果产生任何影响,这也从侧面反映了引入区分性训练的必要性.但是区分性训练相对 MIE 准则来说并不稳定,随着迭代次数的增加,识别率反而会降低.这是由于虽然加大了竞争模型与误识词之间的距离,但是可能造成误识数据过训练,从而使该误识词向其他原来并不是竞争模型的模型靠近,造成新的误识.采用 4 次迭代

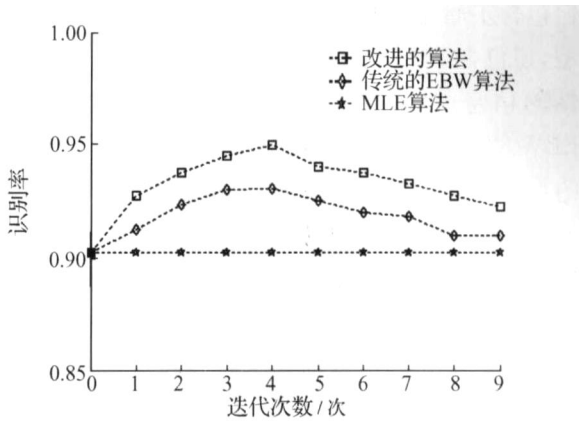


图 2　迭代次数的实验

Fig. 2　The experiment of iterations

一般来说会获得比较理想的结果.这里 h 值取为 1.7,识别率达到 92.5 %.

3.3　实验结果

实验对 MLE 算法,传统的 EBW 算法,以及文中采用的改进的算法进行比较.对于改进的算法,进行 2 组测试.第 1 组用上文提到的易混集构建方法而不加入任何主观经验因素,用 NEW 表示.第 2 组在易混集构建的基础上加入主观的经验因素,用 EXP 表示.分别在注册集和非注册集上对 6 位手语老师中的 5 位给出识别结果.另一位老师的 2 遍数据作为构建训练集的基础,不参与测试.结果如下:

表 1　注册集识别结果

Table 1　Recognition results for registered sets

Signer	MLE	EBW	NEW	EXP
ljh	90.75	91.25	93.25	94.75
llq	92.50	93.00	94.00	95.25
lwr	91.50	92.75	94.50	95.00
mwh	90.25	91.25	92.75	95.00
pfz	93.50	94.00	95.00	96.25
Average	91.70	92.45	93.90	95.25

表 2　非注册集识别结果

Table 2　Recognition results for unregistered sets

Signer	MLE	EBW	NEW	EXP
ljh	67.75	68.57	1.25	75
llq	61.25	61.75	68.5	71.25
lwr	65	66.5	69.5	74.25
mwh	65.25	66.75	69	73.75
pfz	69	70.25	73.25	76.5
Average	65.65	66.75	70.3	74.15

通过表 1 和表 2 可以看到,在注册集和非注册上新的改进算法要大大优于传统的 MLE 算法,这是由于改进算法是在 MLE 基础上的再训练,它不

但包含了原有 MLE 已经训练成熟的基于统计的模型,还包含了 MLE 所不具备的这些模型之间相关性的信息. 此外,由于 D 值选择的困难,导致了 EBW 算法相对 MLE 算法的改进并不明显,而改进的算法可以很好地弥补 EBW 算法的缺陷. 此外,引入主观经验后构造的易混集使识别结果达到最优. 平均识别率相比 MLE,在注册集和非注册集上分别提高了 3.55 %和 8.5 %. 因为加入主观经验后,混合集的构造更加体现了训练集中数据之间的特点,并去掉了很多干扰因素. 这个结果应该是区分性训练的最优结果,可以作为进一步研究的参考界限.

4 结束语

区分性训练方法对传统的 MLE 系统是有效的补充. 重新构建后的模型相对于经典的统计概率模型更能体现手语数据和手语模型之间的相关性. 本论文首次将区分性训练应用于手语识别领域,并对其加以改造,取得了显著的效果.

虽然文中构造的易混集得到了良好的识别结果,但是相对于主观经验所构造的易混集,还有很高的提升空间. 这需要更大量的数据作为实验样本,来挖掘模型之间更深层次的相关性. 由此可见,虽然区分性训练可以使数据量不足够大的 MLE 系统性能提高,但反过来,数据的短缺又会影响区分性训练的效果,这是一对矛盾的统一体.

此外,通过对 MLE 系统的不断改进,已经获得了在注册集上十分令人满意的结果. 然而在非注册集上,识别结果还有很大的提升空间. 下一步应该从数据上着手,如利用有限的数据,生成新的非特定人的数据,来扩大训练集的规模等. 这对进一步研究易混集表的构建也是有指导意义的.

参考文献:

[1] WANG Chunli, GAO Wen. Re-sampling for Chinese sign language recognition by genetic algorithm [A]. GW2005[C]. [s.l.],2005.

[2] DENG J W, TSUI H T. A two-step approach based on PaHMM for the recognition of ASL[A]. Proceedings of The Fifth Asian Conference on Computer Vision[C]. Melbourne, Australia, 2002.

[3] BAHLL R, BROWN P F, SOUZA P V, MERCER R L. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[A]. Proc. 1986 Int. Conf. on Acoustics, Speech and Signal Processing [C]. Tokyo, Japan1986.

[4] NORMANDIN Y. An improved MMIE training algorithm for speaker independent[A]. Proc. ICASSP'91 [C]. Toronto, 1991.

[5] SCHLUTER R, MACHEREY W, RULLER B, NEY H. Comparison of discriminative training criteria and optimization methods for speech recognition[J]. Speech Communication, 2001(34):287 - 310.

[6] ZHENG J, BUTZBERER J, FRANCO H. Scandinavia improved maximum mutual information estimation training of continuous density HMMs[J]. Andreas Stolcke Speech Technology and Research Laboratory, 2001, 15(2):25 - 30.

[7] WOODLAND P C, POVEY D. Large scale discriminative training for speech recognition[J]. In Proc. ITRW ASR[C]. ISCA, 2000.

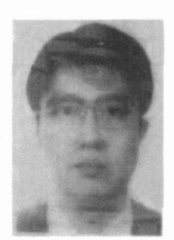
[8] BAHL L R, PADMANABHAN M, NAHAMOO D, GOPALAKRISHNAN P S. An n-best candidates-based discriminative training for speech recognition Applications [J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(1):206 - 216.

[9] CHOW Y L. Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-Best algorithm[A]. Proc. ICASSP'90[C]. Albuquerque, 1990.

作者简介:

王雨轩,男,1980 年生,哈尔滨工业大学硕士研究生,主要研究方向为模式识别、机器学习.

E-mail: yxwang@vilab.hit.edu.cn.



倪训博,男,1978 年生,哈尔滨工业大学博士研究生,主要研究方向为模式识别、机器学习.

E-mail: nixunbo@hit.edu.cn

姜 峰,男,1978 年生,哈尔滨工业大学讲师,主要研究方向为模式识别、机器学习、图像处理、人机交互等.

E-mail: fjiang@hit.edu.cn