基于最大熵方法的统计语言模型

徐延勇　郭忠伟　周献中
（南京理工大学自动控制系，南京 210094）
E-mail：xyy9397@263.net

摘　要　针对现有统计语言模型中存在计算量过大和系统负担过重的问题，该文提出了一种基于最大熵方法的统计语言模型。模型在参数估计阶段，引入约束最优化理论中拉格朗日乘数定理和牛顿迭代算法，以确保模型在多个约束条件中可求出最优化参数值；在特征选择阶段，采用计算近似增益的平行算法，解决模型计算量过大和系统开销问题。将该模型用于汉语句子分析的软件实验中表明：模型具有较高的计算效率和鲁棒性。

关键词　自然语言处理　统计语言模型　最大熵方法

文章编号 1002-8331-(2002)05-0053-03　**文献标识码** A　**中图分类号** TP18

A Statistics Language Model Based on the Maximum Entropy Approach

Xu Yanyong　Guo Zhongwei　Zhou Xianzhong
(Department of automation, Nanjing University of Science and Technology, Nanjing 210094)

Abstract: To solve the problem of computational expensiveness and system spending of the statistics language model in existence, a sort of mathematic algorithms of the statistics language model based on the maximum entropy approach is described detailedly in this article. In parameter estimation stage of the model, Lagrange multipliers from constrained optimization theory and the Newton iterative scaling algorithm are applied to get the optimal parameter values in the more constraint condition; In the feature selection stage, the computing approximate gains in parallel is adopted in order to solve the computational expensiveness of the model and system spending. Software experiments about the Chinese sentence parsing show the model presented in this paper has higher efficiency and robustness.

Keywords: natural language processing, statistics language model, maximum entropy approach

1　引言

目前，基于语料库的统计语言建模方法成为潮流，它通过对语料库进行深层加工、统计和学习，可获得大规模真实语料中的语言学知识以及存在的统计和结构方面的内在规律，并以概率分布的形式描述被观测语句（字符串）中发生某种语言学类别的概率，此类模型的代表主要有[1]：上下文无关模型，N-gram 模型，N-pos 模型等。这些模型由于具有较强的知识表达能力和易于计算等优点，已被广泛应用于汉语歧义分析、词性标注、语音识别等领域。随着应用领域和词汇量的不断扩大，现有统计语言模型在使用过程中逐步暴露出了计算量过大和系统负担过重等问题。针对这一问题，该文提出了一种基于最大熵（Maximum Entropy，ME）方法的统计语言模型。

2　基于最大熵方法的统计语言模型框架

2.1　问题描述

在自然语言处理这一随机过程中，所有最终输出值构成了语言学类别有限集 Y，对于每个 $y\in Y$，其生成均受上下文信息 x 的影响和约束。已知与 y 有关的所有上下文信息组成的集合为 X，则模型的目标是：给定上下文 $x\in X$，计算输出为 $y\in Y$ 的条件概率 $p(y|x)$。如 p（他/认真/学习）≈0.02，p（他/认真/美丽）≈0，词“认真”对后面词的出现有较强的约束力。模型的输入是从经人工标注的训练数据中所抽取的训练样本集 $T=\{(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_N,y_N)\}$，$(x_i,y_i)$ 表示在语料库中出现的第 i 种语言学类别，这些类别都已有确定的取值 y_i 及其对应的上下文环境 x_i。可以用概率分布的极大似然对训练样本表示如下：

$$\tilde{p}(x,y)=\frac{freq(x,y)}{N} \tag{1}$$

$freq(x,y)$ 是 (x,y) 在 T 中出现的次数。

2.2　上下文谓词、特征和约束

y 的生成只与其上下文 x 的部分信息有关，从 x 中找出对 y 的取值有用的知识才是模型所追求的目标，这有用的部分知识也就是 ME 方法所要寻找的特征。在此引入了上下文谓词、特征及其约束。

（1）上下文谓词用 $cp:Y\rightarrow\{ture,false\}$ 表示，以检测上下文 $y\in Y$ 中有用信息的存在和缺少，根据情况返回 *true* 或者 *false*。

（2）特征 f 是关于 (x,y) 的二值表征函数，把上下文谓词用在特征中，则特征函数表示如下：

$$f(x,y)=\begin{cases}1 & \text{if } cp(y)=\text{true } \&\ \&x'=x\\ 0 & \text{否则}\end{cases} \tag{2}$$

基金项目：国防科工委跨行业基金项目资助

作者简介：徐延勇，博士生，主要从事自然语言处理的研究工作。郭忠伟，博士生，主要从事自然语言生成的研究工作。周献中，教授、博士生导师，主要从事信息系统工程、指挥自动化、辅助决策和计算语言学的研究工作。

ME 方法要选的特征必须能较完整地表达训练语料中的数据，对模型真正有用。但与模型有关的候选特征 F 组成的集合很大。由此有约束：

(3)设 $\tilde{p}(f)$ 为特征 f 对于经验概率分布 $\tilde{p}(x,y)$ 的数学期望，表示为

$$\tilde{p}(f)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{3}$$

$p(f)$ 为特征 f 对于由模型确定的概率 $p(x,y)$ 的数学期望，表示为

$$p(f)=\sum_{x,y}p(x,y)f(x,y) \tag{4}$$

而 $p(x,y)=p(x)p(y|x)$，令 $p(x)=\tilde{p}(x)$，$\tilde{p}(x)$ 是 x 在训练样本中的观测概率。则限定所求模型的概率为在样本中观察到事件的概率，而不是所有可能事件的概率。若 f 对模型有用，则约束为：

$$p(f)=\tilde{p}(f) \tag{5}$$

2.3 最大熵方法的引入

在相同训练数据的基础上，这一节介绍了 ME 方法[3,4]和最大似然估计 MLE(Maximum Likelihood Estimation)两种方法如何把上面所示的局部知识有效结合起来，并对两种方法进行了比较。

(1)假设存在 k 个特征 $f_i(i=1,2,...,k)$，属于约束所产生的模型集合空间为：

$$C=\{p\in P|p(f)=\tilde{p}(f)\},i\in\{i=1,2,...,k\} \tag{6}$$

满足约束条件的模型有很多，模型的目标是产生在约束集下具有最均匀分布的模型，用条件熵作为条件概率 $p(y|x)$ 均匀性的一种数学测量方法。

$$H(p)=-\sum_{x,y}\tilde{p}(x)p(y|x)\log p(y|x) \tag{7}$$

这里 $0\leqq H(p)\leqq\log|y|$

ME 方法提出了一个约束最优化问题：在与约束集合 C 一致的模型中，选择具有最大熵的 $p^*\in C$。在有用特征 f 的基础上进行推论，它能产生最优化和唯一无偏估计值 p^*。

$$p^*=\arg\max_{p\in C}\ H(p) \tag{8}$$

特征 f_i 的权重用相对应的参数 λ_i 表示，则满足最大熵的条件概率 $p(y|x)$ 用指数形式表示为：

$$p_\lambda(y|x)=\frac{1}{Z_\lambda(x)}\exp\left(\sum_i\lambda_i f_i(x,y)\right)$$

$$Z_\lambda(x)=\sum_y\exp\left(\sum_i\lambda_i f_i(x,y)\right) \tag{9}$$

$Z_\lambda(x)$ 是保证对所有 x 使得 $\sum_y p_\lambda(y|x)=1$ 归一化常量。而在相同训练数据基础上的 MLE 方法则如(2)表示：

(2)令 Q 是(9)式指数形式的模型集合，$L(p)$ 是训练样本中的条件概率，则 p^* 是唯一的，即

$$Q=\left\{p|p_\lambda(y|x)=\frac{1}{Z_\lambda(x)}\exp\left(\sum_i\lambda_i f_i(x,y)\right)\right\} \tag{10}$$

$$L(p)=\sum_{a,b}\tilde{p}(x,y)\log p(y|x)$$

$$p_*=\arg\max_{q\in Q}\ L(p)$$

(3)ME 与 MLE 是建立统计语言模型的两种不同方法，但在训练数据一致的模型中可产生相同的结论。即如果 $P^*\in P\cap Q$，则有 $P^*=argmax_{q\in Q}L(q)=argmax_{p\in P}H(p)$，$P^*$ 是唯一的。在 ME 方法里，P^* 不允许任何超过约束的额外假定；而在 MLE 方法中，P^* 尽可能与训练数据密切联系。

3 基于最大熵方法的统计语言模型构建

利用 ME 方法建立统计语言模型的过程分为两步：参数估计和特征选择。特征选择的任务是选出对模型有表征意义的特征；参数估计是利用 ME 原理对每一个特征进行参数估值，使每一个参数与一个特征相对应，以此建立所求模型。

3.1 参数估计

当约束条件只有一个时，能通过分析方法直接得到 P^*；但当约束条件很多，就需要用数值计算的方法得到。如何从多约束条件中取得最优化的参数值 λ^* 和模型 p_{λ^*}，这里，从约束最优化理论中引入了拉格朗日乘数定理，具体步骤如下：

步骤 1 用固有约束最优化问题作为初始问题

$$p^*=\arg\max_{p\in C}\ H(p) \tag{11}$$

步骤 2 设参数 λ_i 为拉格朗日因子，定义拉氏函数：

$$\wedge(p,\lambda)=H(p)+\sum_i\lambda_i(p(f_i)-\tilde{p}(f_i)) \tag{12}$$

步骤 3 保持 λ 固定，在所有 $p\in P$ 中计算 $\wedge(p,\lambda)$ 的无约束最大值以及对偶函数 $\Psi(\lambda)$：

$$p_\lambda=\arg\max_{p\in C}\wedge(p,\lambda) \tag{13}$$

$$\Psi(\lambda)=\wedge(p_\lambda,\lambda)=-\sum_x\tilde{p}(x)\log Z_\lambda(x)+\sum_i\lambda_i\tilde{p}(f_i) \tag{14}$$

步骤 4 形成无约束对偶最优化问题

$$\lambda^*=\arg\max\ \Psi(\lambda) \tag{15}$$

该定理的 Kuhn-Tucker 准则：在合适条件下，初始问题和对偶问题是紧密联系的。设服从约束集合 C 的 ME 方法有(2-9)的参数形式 p_{λ^*}，如果 λ^* 是对偶问题的解，那么 p_{λ^*} 就是初始问题的解，$p_{\lambda^*}=p^*$。即能求出对偶函数 $\Psi(\lambda)$ 中最大 λ^* 值的算法同样也能用于求解 $p\in C$ 中 $H(p)$ 的最大值 P^*，这样，就较好地解决了多约束条件带来的求解问题。同时，从数字优化的观点来看，函数 $\Psi(\lambda)$ 是平滑和严格凸-的。可以用很多种数学方法来计算 λ^*，包括有坐标爬升法，梯度爬升法和共轭梯度法。这里，对 ME 方法特别设计了一种迭代算法 (Iterative Scaling Algorithm，ISA)算法来求解，要求：对所有的 $i,x,y,f_i(x,y)\geq 0$，这明显成立，因为特征函数是二值函数；对所有的 x,y，特征函数都满足 $\sum_i f_i(x,y)=1$。算法步骤如下：

输入：特征函数 $f_1,f_2,...,f_n$；经验概率 $\tilde{p}(x,y)$

输出：最优化参数值 λ_i^*；最优化模型 p_{λ^*}

步骤 1 设 $\lambda_i=0$ 对所有的 $i\in\{1,2,...,n\}$

步骤 2 对每个 $i\in\{1,2,...,n\}$：

a. 设增量 $\Delta\lambda_i$ 是如下方程的解

$$\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y)\exp(\Delta\lambda_i f^\#(x,y))=\tilde{p}(f_i)$$

$$f^\#(x,y)=\sum_{i=1}^n f_i(x,y)$$

b. 根据 $\lambda_i+\Delta\lambda\rightarrow\lambda$ 更新 λ_i 的值

步骤 3　如果不是所有的 λ_i 都收敛，转步骤 2；当所有的 λ_i 都收敛时，结束。

如果对所有的 $x,y,f^{\#}(x,y)=M$（常量），则 $\Delta\lambda_i$ 由下式直接得出：

$$\Delta\lambda_i=\frac{1}{M}\log\frac{\tilde{p}(f_i)}{p_\lambda(f_i)} \tag{16}$$

如果 $f^{\#}(x,y)$ 不是常量，$\Delta\lambda_i$ 必须通过数字计算。一种简单而有效的方法是用牛顿法计算，选择合适的 a_0 以及相配定义域 g，它通过下面的循环迭代求方程式 $g(a_*)=0$ 的解：

$$a_{n+1}=a_n-\frac{g(a_n)}{g'(a_n)} \tag{17}$$

设 N 是训练样本的数量，A 是预测的数量，V 是对一个给定事件 (x_i,y_i) 有效特征的平均数量，则每步迭代运行时间为 $O(N|A|V)$。

3.2　特征选择

ME 方法不直接与特征选择联系起来，它仅仅提供了一种如何结合模型中多约束的巧妙方法。文章提出了一种在大数量可能约束中自动选择特征的数学方法，并且提供了一系列优化算法来解决计算量过大的问题。

3.2.1　特征选择的问题描述

在候选特征集合 F 中，只有一小部分子集即有效特征集合 S 能用在最后的模型里。S 要尽可能抓住自然语言处理的随机过程中其期望值能被可靠估计的特征，用 $C(S)$ 定义有效特征集 S 决定的模型，$C(S)$ 中只有一部分能满足等式 $\tilde{p}(f)=p(f)$。

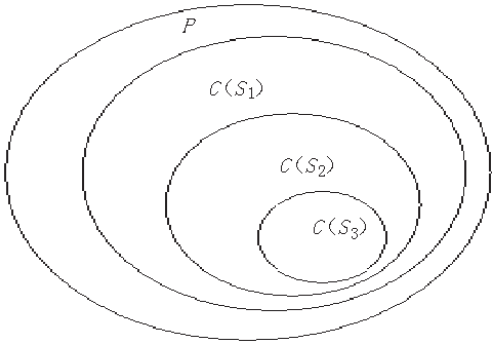
为了计算 S，对特征采用增量方法：连续对 S 增加特征来增大 S，每步中增加的特征由训练数据决定。每次对 S 增加一个候选特征 $\hat{f}$ 时，相应的一个线形约束也强加在模型空间 $C(S)$ 上。这样，使得 $C(S)$ 收缩，用 P 的一系列嵌套子集表示这个可能模型空间的缩小（图 1 所示）；并且得到一个新的有效特征集 $S\cup\hat{f}$，用 $C(S\cup\hat{f})$ 定义新的模型集合空间为：

$$C(S\cup\hat{f})=\{p\in P|p(f)=\tilde{p}(f)\ \ f\in S\cup\hat{f}\} \tag{18}$$

在这个模型集合空间中，最优化模型为：

$$p_{S\cup\hat{f}}=\arg\max_{p\in c(S\cup\hat{f})}H(p) \tag{19}$$

有最大熵的模型 $p_{S\cup\hat{f}}$ 反应了不断增加的知识，它能成为随机过程的更精确描述。



$C(S_1)\supset C(S_1)\supset C(S_1)\cdots S_1\subset S_2\subset S_3\cdots$

图 1

3.2.2　特征选择计算

以训练数据的对数似然作为特征选择依据，用 $L(p_s)$ 表示由 S 决定模型的对数似然。则建立模型的目标是要选出使（20）式增加最多的特征 $\hat{f}$。

$$\Delta L(S,\hat{f})=L(p_{S\cup\hat{f}})-L(p_s) \tag{20}$$

这里，$L(p_s)=\sum_{x,y}\tilde{p}(x,y)\log(y|x)$

如果用该算法，每选一个特征都需要对所有的候选特征调用 ISA 算法，对参数 λ 重新计算，并且要对训练数据的对数似然进行计算，然后选出一个使模型的对数似然 $\Delta L(S,\hat{f})$ 增加最多的特征，计算量很大，这几乎是不可操作的。

该文提出了一种用相似度增益 $\sim\Delta L(S,\hat{f})$ 来代替 $\Delta L(S,\hat{f})$ 的算法：在 p_s 中包含有对应 S 中每个特征的参数集合；在模型 $p_{S\cup\hat{f}}$ 中也包含这个集合，另外还增加一个新参数 a 来对应增加的特征 $\hat{f}$。但当一个特征 $\hat{f}$ 加入 S，一个新的约束被引入时，所有参数的最优化值都发生了改变。为了使参数再排序计算变得容易，要使特征 $\hat{f}$ 的增加仅仅影响 a 的值，而对其他相关的 λ 值没有影响，则包含特征 $S\cup\hat{f}$ 的最好模型有如下形式：

$$p^{a}_{S,f}=\frac{1}{Z_a(x)}p_s(y|x)e^{af(x,y)} \tag{21}$$

这里 $Z_a(x)=\sum_y p_s(y|x)e^{af(x,y)}$

最大近似增益的计算可表述为：

$$\sim\Delta L(S,\hat{f})=\max_a G_{s,f(a)} \tag{22}$$

$$G_{s,f(a)}=L(p^{a}_{S,f})-L(p_S)=-\sum_x\tilde{p}(x)\log Z_a(x)+a\tilde{p}(f)$$

相应的最优化模型为：

$$\sim p_{S\cup f}=\arg\max_{p^{a}_{S,f}} G_{s,f(a)} \tag{23}$$

在实际应用软件中，训练数据不能完全存储在内存而需要从硬盘读取数据，这将产生很大的时间花费。为了避免在每步迭代中遍历整个训练样本中每个事件所代来的计算浪费，在计算 $\sim\Delta L(S,\hat{f})$ 时，作者特别应用了一种只需要遍历训练数据的很少次数的算法：计算近似增益的平行算法（Computing Approximate Gains in Parallel，CAGIP）：

输入：候选特征集合 F，经验概率 $\tilde{p}(x,y)$，初始模型 p_S

输出：每个候选特征 $f\in F$ 的近似增益 $\sim\Delta L(S,\hat{f})$

步骤 1　计算在训练数据中每个特征 $f\in F$ 的期望值 $\tilde{p}(f)$

步骤 2　对每个 x，定义对 x 有效的特征 f 的集合 $F(x)\subseteq F$

$F(x)=\{f\in F|f(x,y)p_S(y|x)\tilde{p}(x)>0$

步骤 3　对每个特征 f，使得

$$r(f)=\begin{cases}1 & \tilde{p}(f)\le p_S(f)\\ -1 & \text{其他}\end{cases}$$

步骤 4　对每个特征 $f\in F$，初始化每个特征 $a(f)\to 0$

步骤 5　对每个特征 $f\in F$，重复下面的步骤直到 $a(f)$ 收敛

(a) 对每个特征 $f\in F$ 设

$G'(f)\to\tilde{p}(f)$

$G''\to 0$

（下转 121 页）

全局管脚： 1/6 (16%)
I/O 管脚： 104/464 (22%)
逻辑单元： 1825/9984 (18%)
嵌入单元： 8/384 (2%)
EAB： 1/24 (4%)
输入： 6066/39936 (15%)
输入管脚占用数：62
输出管脚占用数：43
寄存器数： 216

整个电路的工作主频达到了 30Mhz，并且为了方便测试仪的测试引出脚工作，将所有的管脚均布在了芯片的外围两圈，为以后的工作也提供了方便。

5 结束语

随着 VLSI 工艺的不断提高，CPLD 芯片的规模也越来越大，其研制开发费用较低，设计周期短，不需要电路设计人员具有深层次的 IC 知识，且有利于对知识产权的保护。因此受到世界范围内电子工程设计人员的广泛关注和普遍欢迎。该文给出了基于 MIL-STD-1750A 的嵌入式微处理器的浮点执行部件的 CPLD 的实现，所得的结果达到了预期的要求。

（收稿日期：2000 年 11 月）

参考文献

1.十六位计算机指令系统结构.国防科学技术工业委员会
2.PACE1750A，MIL_STD_1750A MICROPROCESSOR INSTRUCTION Reference[M].Performance Semiconductor Corp
3.An American National Standard IEEE Standard for Binary Floating_ Point Arithmetic[S].American National Standard Institute，Approved 1985
4.白中英.计算机组成原理[M].科学出版社，1994
5.宋万杰，罗丰，吴顺君.CPLD 技术及其应用[M].西安电子科技大学出版社，1999
6.薛宏熙，边计年，苏明.数字系统设计自动化[M].清华大学出版社，1995

（上接 55 页）

(b)对每个 x 的每个 $f\in F(x)$，通过下式更新 $G'(f)$ 和 $G''(f)$

$$G'(f)-\tilde{p}(x)p^{a}_{S,f}(f|x)\rightarrow G'(f)$$

$$G''(f)-\tilde{p}(x)p^{a}_{S,f}((f-p^{a}_{S,f}(f|x))^2|x)\rightarrow G''(f)$$

这里 $p^{a}_{S,f}(f|x)=\sum_y p^{a}_{S,f}(y|x)f(x,y)$

(c)对每个特征 $f\in F$，通过如下更新 $a(f)$

$$a(f)+\frac{1}{r(f)}\log\left(1-\frac{1}{r(f)}\frac{G'(f)}{G''(f)}\right)\rightarrow a(f)$$

步骤 6 把 $a(f)$ 代入(22)中，计算 $\sim\Delta L(S,\hat{f})$

经过步骤 5 的迭代后，每个候选特征 f 的 $a(f)$ 值接近于最优化值 $a^*(f)$，并且增益 $G_{S,f}$ 与最大相似增益 $\sim\Delta L(S,\hat{f})$ 接近。有这样的准备后，就可以进行特征选择了，相应的算法如下：

输入：候选特征集合 F 经验概率 $\tilde{p}(x,y)$
输出：有效特征集合 S 结合这些特征的模型 P_S
步骤 1 初始化 $S=0$
步骤 2 对每个候选特征 $f\in F$：
a 用 ISA 计算模型 $p_{S\cup\hat{f}}$
b 用 CAGIP 计算 $\sim\Delta L(S,\hat{f})$
步骤 3 检查终止条件，如终止，结束；否则，继续
步骤 4 选择有最大 $\sim\Delta L(S,\hat{f})$ 的特征 $\hat{f}$
步骤 5 把 $\hat{f}$ 加入 S
步骤 6 用 ISA 计算 p_S
步骤 7 返回步骤 2

终止条件是当所有的有用特征都被选出时结束。一个可行的标准是在提供的训练样本数据中，增加的特征在似然估计上没有增量时，则认为这个特征不能预测当前随机事件中的本质信息而舍去。

4 实验及结论

笔者已将文章所述的模型与算法应用于汉语句子分析，利用 VC++ 6.0 平台自主开发了相关软件。该软件在进行汉语句子分析时，将 2 000 万字已经进行了人工标注的《人民日报》语料库作为实验的训练数据，用随机方式从人民日报中抽取句子经分词后作为测试句子输入进行实验，实验结果表明：基于最大熵模型的统计语言模型对汉语句子的分析准确率达到 90.1%，召回率达到 89.2%；分析测试句子的耗时长度与句长成线形比例关系。详细实验结果将另有论文专述。

与其它统计语言学建模方法相比，ME 方法提供了正确结合模型化数据中有用知识的一种巧妙办法。它用特征的形式表示自然语言处理中各种语言学类别知识并与其自身实现关联，其特征没有额外的独立假定或内在约束；一旦实验者发现新的特征，不用修改数学模型，就能够给分析模型添加任意不同或复杂的特征集，这使得模型应用在不同领域时的可移植性强；并且实验者只需发现什么特征是有用的，而不是如何去运用这些特征，减少了过程中的人为因素。（收稿日期：2002 年 1 月）

参考文献

1.关毅，张凯，付国宏.基于统计的计算语言模型[J].计算机应用研究，1999；(6)：26-28
2.Darroch J N，Ratcliff D.Generalized iterative scaling for log-linear models[J].The Annals of Mathematical Statistics，1972；43(5)：1470-1480
3.Au R Rosenfeld.Adaptive language modeling using the maximum entropy principle[C].In：Proceedings of the Human Language Technology Workshop，ARPA：1993：108-113
4.Rosenfeld R.A maximum entropy approach to adaptive statistical language modeling[J].Computer，Speech，and Language，1996；10
5.Jaynes E T.Notes on present status and future prospects[C].In：Grandy W T，Schick L H eds.Maximum Entropy and Bayesian Methods，Kluwer：1990：1-13
6.穗志方，赵军，愈士汶.统计句法分析建模中基于信息论的特征类型分析[J].计算机学报，2001；(2)：144-151