语言模型中一种改进的最大熵方法及其应用*

李涓子 黄昌宁

(清华大学计算机科学与技术系 北京 100084)

(清华大学智能技术与系统国家重点实验室 北京 100084)

E-mail: ljz@s1000e.cs.tsinghua.edu.cn

摘要 最大熵方法是建立统计语言模型的一种有效的方法,具有较强的知识表达能力.但是,在用现有的最大熵方法建立统计模型时存在计算量大的问题.针对这一问题,提出了一种改进的最大熵方法.该方法使用互信息的概念,通过Z-测试进行特征选择.将该方法应用于汉语的义项排歧中,实验表明,该算法具有较高的计算效率和正确率.

关键词 语言模型,最大熵模型,参数估计,特征选择,互信息,Z-测试.

中图法分类号 TP18

语言模型试图反映、记录并使用自然语言中存在的规律[1].近几年在自然语言处理的研究过程中发现,最大熵方法是一种建立统计语言模型的有效方法,具有较强的知识表达能力.但是,在用现有的最大熵方法建立统计模型时存在计算量过大的问题[2].本文针对这一问题,提出了一种新的特征选择算法.算法使用互信息的概念,通过Z-测试的方法进行特征选择.在特征选择过程中,用新建模型与引用模型的Kullback-Leibler距离来调整所选出的特征.实验表明,这种算法具有较高的计算效率.

本文首先叙述最大熵原理,介绍已有的参数估计和特征选择方法,并对其进行评价;然后给出使用Z-测试的特征选择算法,最后将这种改进的最大熵方法应用于汉语的义类排歧中.

1 最大熵原理

最大熵原理最初是由E. T. Jayness在1950年提出的,Della Pietra等人于1992年首次将它应用于自然语言处理的语言模型建立中[1].本文只是简单介绍最大熵原理,更详细的叙述请见参考文献[3,4].

直觉上讲,最大熵原理的基本思想是:给定训练数据即训练样本,选择一个与所有的训练数据一致的模型.比如在英语中,对于一个具有词性歧义的词条,如果发现一个名词前为一个冠词的概率为50%,而在名词前为一个形容词的概率为30%,则最大熵模型应选择与这些观察一致的概率分布.而对于除此之外的情况,模型赋予的概率分布为均匀分布.

1.1 问题描述

设随机过程P所有的输出值构成有限集Y,对于每个输出$y\in Y$,其生成均受上下文信息x的影响.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,即对$p(y|x)$进行估计.$p(y|x)$表示在上下文为x时,模型输出为y的条件概率,其中$y\in Y$且$x\in X$.如:对于义类歧义问题,集合Y是具有义类歧义的某一词W的所有可能义类组成的集合,集合X为对词W的每次出现,为其选定的上下文环境所组成的集合.

---

* 本文研究得到国家自然科学基金重点项目资助.作者**李涓子**,女,1964年生,博士生,主要研究领域为计算机语言学.**黄昌宁**,1937年生,教授,博士生导师,主要研究领域为计算语言学,人工智能.

本文通讯联系人:李涓子,北京100084,清华大学计算机科学与技术系

本文1997-12-11收到原稿,1998-03-12收到修改稿

1.2 训练数据

模型输入是经过人工排歧或从已标注过的语料库中抽取出的大量(x,y)训练样本,即对在语料库中有歧义的对象的每次出现,都已有确定的取值y及其对应的上下文环境x.可以用概率分布的极大似然对训练样本进行表示.即

$$\tilde{p}(x,y)\equiv\frac{freq(x,y)}{\sum_{x,y}freq(x,y)}, \tag{1}$$

其中$freq(x,y)$是(x,y)在样本中出现的次数.

1.3 特征、特征函数及约束

由问题描述可知,随机过程P与上下文信息x有关,但如果考虑所有与y同现的上下文信息,建立的模型会很繁琐,而且从语言知识上来讲,y的生成只与其上下文中的部分信息有关.因此,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.而这些有用的知识正是最大熵模型所要寻找的特征.

定义1. 特征

设$x\in X$且$x=w_1w_2\ldots w_n$,而c是x的一个子串(长度$\geqslant 1$),若c对$y\in Y$具有表征作用,则称(c,y)为模型的一个特征.

特征分为原子特征和复合特征.若串c的长度为1,则称(c,y)为原子特征,否则,称(c,y)为复合特征.

定义2. 特征函数

特征函数是一个二值表征函数,表示(x',y')是否与特征(c,y)有关,定义(x',y')关于特征(c,y)的特征函数为

$$f_{(c,y)}(x',y')=\begin{cases}1 & \text{若 } c \text{ 是 } x' \text{ 的子串,且 } y'=y\\ 0 & \text{否则}\end{cases}. \tag{2}$$

由以上定义可以看出,样本中出现在歧义对象周围的所有的词和该对象的确定值一起都可以作为模型的特征,因此,与模型有关的候选特征组成的集合会很大.但模型选出的特征只是真正对模型有用的特征,是候选特征集合的一个子集,它能较完整地表达训练语料中的数据.由此引入约束.

定义3. 约束

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$p(x)=\tilde{p}(x)$,则限定所求模型的概率为在样本中观察到事件的概率,而不是所有可能出现的事件的概率.若f对模型有用,则令

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

称式(5)为约束.

1.4 最大熵原理

假设存在n个特征$f_i(i=1,2,\ldots,n)$,则模型属于约束所产生的模型集合,即

$$C=\{p\in P\mid p(f_i)=\tilde{p}(f_i),i\in\{1,2,\ldots,n\}\}, \tag{6}$$

而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,而条件概率$p(y|x)$均匀性的一种数学测量方法为条件熵,定义为

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

其中$0\leqslant H(p)\leqslant\log|y|$.

最大熵原理. 若在允许的概率分布C中选择模型,具有最大熵的模型$p_*\in C$即为所选模型.即

$$p_*=\arg\max_{p}H(p). \tag{8}$$

2 参数估计及特征选择

利用最大熵建立语言模型的过程分为两步:特征选择和参数估计.特征选择的任务是选出对模型有表征意义的特征;参数估计用最大熵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.

2.1 参数估计

Danroch 和 Ratcliff 于 1972 年提出一个称为 GIS(generalized iterative scaling algorithm)的算法,该算法是一般的迭代算法.Della Pietra 等人于 1995 年根据所处理的问题对算法作了进一步改进,提出了 IIS(improved iterative scaling algorithm)算法,算法设满足最大熵条件的概率 $p(x,y)$ 具有 Gibbs 分布的形式

$$p(y|x)=\frac{1}{Z_\lambda(x)}e^{\sum_i \lambda_i f_i(x,y)}, \tag{9}$$

其中

$$Z_\lambda(x)=\sum_y e^{\sum_i \lambda_i f_i(x,y)}, \tag{10}$$

$Z_\lambda(x)$ 为归一常量,保证对所有 x,$\sum_y p_\lambda(y|x)=1$.

2.2 特征选择

无论 GIS 还是 IIS 的参数估计方法提供的均是求解 λ 值的方法,保证以 λ 建立的模型不含有任何额外的假设.这两个算法并不能保证模型所含特征是具有良好表征意义的特征,因此,在建立模型中十分重要的一部分工作是特征选择.Della Pietra 等人提出的原子特征算法思想是:开始设特征集 S 为空,此后不断向 S 中增加特征,每次增加的特征由训练数据决定.以训练数据的对数似然作为特征选择的依据,即若 S 为已选中的特征集,$\tilde{f}$ 为候选特征,用 $L(p_s)$ 表示由 S 决定的模型的对数似然,则每次选出的 $\tilde{f}$ 应该为使公式

$$\Delta L(s,\tilde{f})\equiv L(p_{s\cup\tilde{f}})-L(P_s), \tag{11}$$

增加最多的特征.其中

$$L(P_s)=\sum_{x,y}\tilde{p}(x,y)\log(y|x). \tag{12}$$

该算法的致命弱点是计算量大,每选一个特征都需要对所有的候选特征调用 IIS 算法,对 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增加最多的特征,这几乎是不可操作的.为了使特征选择过程可行,Della Pietra 等人又给出了一系列优化算法,如在向模型中加入一个新的特征时,保持前面 IIS 过程估计的 λ 值不变,只用 IIS 计算新加入特征的对应值,当所有特征选出后,重新调用一次 IIS 过程,对所有 λ 进行一次重新计算,这种方法虽然可以加快特征选择的过程,但不能保证每次加入模型的特征是最好的.

3 使用 Z-测试的特征选择算法

建立最大熵模型的关键是要选出具有预期作用的特征,只有这样才能保证得到的解是对模型最有用的解.虽然 Della Pietra 等人的原子特征选择方法,可以选出最好的有预期作用的特征,但这种方法完全建立在数学运算的基础之上,存在着计算量大的问题.

既然特征选择的目的是要选出对模型具有预期作用的上下文信息,则这个信息与所要预期的值具有较密切的搭配关系.本文正是从这一假设出发,提出一种使用互信息概念,采用 Z-测试的方法来进行特征选择的算法.

3.1 原子特征选择的问题描述

已知训练数据中的 N 个训练样本 $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$,其中 $x_i\in X$ 且 $y_i\in Y$,设 $A=\{a_j|a_j$ 是 x_i 的子串,且 a_j 的长度为 $1,i=1,2,\ldots,N\}$,则原子特征选择的意义是从 A 中选出能够充分表征 Y 的不同取值的最小特征集 $F=\{f_1,f_2,\ldots,f_n\}$,其中 $f_i=(a_i,y_i)(a_i\in A,y_i\in Y,i=1,2,\ldots,n)$ 为原子特征.

在此,我们采用 Kullback-Leibler 距离来测定特征所确定模型的质量.设 $\tilde{p}$ 是由训练语料确定的概率模型,p 为由特征集确定的模型,则 Kullback-Leibler 距离定义为

$$D(\tilde{p} \| p)=\sum_{x,y}\tilde{p}(x,y)\log\frac{\tilde{p}(y|x)}{p(y|x)}. \tag{13}$$

最终要找的模型 p 为

$$p=\min_{p}D(\tilde{p} \| p). \tag{14}$$

3.2 利用 Z-测试进行原子特征选择的依据

(1) 互信息[5]可衡量搭配的强度

特征选择的目的是要选出对模型具有预期作用的上下文信息,所以这个信息应与所预期的值具有较密切的搭配关系,而信息论中的互信息正是测量搭配强度的一个物理量.对应于我们要解决的问题为:若某一上下文信息对 y 有表征意义,则 y 与该上下文的互信息较大.

(2) Z-测试[6]可作为互信息的一个测度

虽然互信息可以作为描述搭配强度的物理量,但是,如果特征选择直接确定选择互信息大于某一阈值的上下文信息为特征时,则对不同互信息的分布,设定的阈值也不相同,这样,算法难以操作.而 Z-测试可以将互信息的分布进行标准变换,将其变为标准的正态分布,这样,不论互信息如何分布,都可以从一个统一的阈值开始进行求解.

3.3 原子特征选择算法

输入:训练样本 $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$;

输出:特征矩阵 $D_{m\times n}$.其中

$m=|Y|$,即 y 的所有可能取值的个数;

$n=|\{a_j|a_j$ 是 x_i 的子串,且 a_j 的长度为 $1,i=1,2,\ldots,N\}|$,即与 y 的不同值同现的候选特征集合中的元素个数.

$$d_{ij}=\begin{cases}1 & a_j \text{ 为 } y_i \text{ 有表征作用}\\ 0 & \text{否则}\end{cases}.$$

过程

步骤 1.

· 由样本 $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ 得到 Y 中元素与 A 中元素的同现次数矩阵 $F_{m\times n}$,其中

$$f_{ij}=f(y_i,a_j)=|\{(x_k,y_k)|a_j \text{ 是 } x_t \text{ 的子串,且长度为 } 1,y_k=y_i,s=1,2,\ldots,N\}|.$$

· 计算互信息矩阵 $I_{m\times n}$,其中

$$I(i,j)=\log_2\frac{p(y_i,a_j)}{p(y_i)p(a_j)}=\log\frac{Mf_{ij}}{f(y_i)f(a_j)},$$

其中 M 为语料库的大小,而 $f(y_i)$ 和 $f(a_j)$ 分别为 y_i 和 a_j 在语料库中出现的次数.由互信息的定义可知:当 $I_{ij}\gg 0$ 时,y_i 与 a_j 完全并列,因此,(a_j,y_i) 可作为模型的特征.

步骤 2.

对每个 y_i,

· 计算 y_i 的互信息均值

$$E_i=\frac{1}{n}\sum_{j=1}^{n}I(y_i,a_j); \tag{15}$$

· 计算其均方差

$$u_i=\frac{1}{n}\sum_{j=1}^{n}(I(y_i,a_j)-E_i)^2. \tag{16}$$

步骤 3.

用 Z-测试对每个 y_i 生成表征向量 $(d_{i1},d_{i2},\ldots,d_{im})$;

· 对每个 $y_i(i=1,2,\ldots,n)$,

* 对每个 a_j,计算

$$z_{ij}=\frac{I(y_i,a_j)-E_i}{\sqrt{u_i}}, \tag{17}$$

* 若 $z_{ij}>T$，则 a_j 与 y_i 为选出的一个特征，令 $d_{ij}=1$；否则 $d_{ij}=0$；

步骤 4.

· 用选出的原子特征集合 $S=\{f_1,f_2,\ldots,f_k\}$ 调用 IIS 算法，得到 $(Z,\lambda_1,\lambda_2,\ldots,\lambda_k)$；

· 用公式(9)和(10)计算 $p(y|x)$.

步骤 5.

· 计算由 $p(y|x)$ 确定的模型与经验概率分布模型 $\tilde{P}(y|x)$ 的距离 $D(p\|\tilde{p})$；

· 用 $D(p\|\tilde{p})$ 与上次的 $D'(p\|\tilde{p})$ 比较：若 $D-D'<\varepsilon$，则过程结束；否则，$T=T-\Delta T$，转步骤 3.

3.4　阈值 T 的确定

(1) T 初值的确定

从算法可以看出，在经过式(17)的运算后，已将互信息的分布变为正态分布. 从概率论可知：正态分布在区间[−3，+3]内，其整个概率覆盖度可达 99%左右. 因此，T 可以在[−3，+3]内进行取值. 因为开始时要选出表征意义大的特征，所以应赋予 T 一个较大的初值.

(2) T 阈值的变化

初值确定后，以后每次以一个步长 ΔT 减少，这就意味着每次根据 T 选中的特征不是一个，而是具有同等表达程度的一个候选特征子集，且选出的子集中包含上一次选出的特征集合. 因此，在进行下一次的参数估计时，对于以前的特征其初值可以从上次确定的值开始，这样做可以节省大量运算时间. 特征选择过程最终得到的特征集合是它所确定的模型的 $D(p\|\tilde{p})$ 较小且具有较一般表征意义的集合.

3.5　两个原子特征选择算法的计算量比较

(1) Della Pietra 的特征选择算法的计算量分析

该特征选择算法每次确定一个特征时的计算量由两部分组成，即调用 IIS 对每一候选特征进行参数估计和计算模型的对数似然. 总的计算量可表示为

$$C_1=n*(IIS_1+L_1)+(n-1)*(IIS_2+L_2)+\ldots+(n-k)*(IIS_k+L_k). \tag{18}$$

其中 n 为候选特征集合中候选特征的个数，k 为最终特征集合中的特征个数，IIS_i 和 L_i 分别表示在选第 i 个特征时参数估计的计算量和对数似然的计算量.

设 $IIS_{\min}$ 为在 k 次特征选择过程中，参数估计过程所需的最少时间，则

$$C_1\geqslant n*(IIS_{\min}+L_1)+(n-1)*(IIS_{\min}+L_2)+\ldots+(n-k)*(IIS_{\min}+L_{\min}).$$

由公式(12)可知，$L_1=L_1=\ldots=L_k$，则

$$\begin{aligned}C_1&\geqslant n*(IIS_{\min}+L_1)+(n-1)*(IIS_{\min}+L_1)+\ldots+(n-k)*(IIS_{\min}+L_1).\\&=\frac{(k+1)(2n-k)}{2}\cdot IIS_{\min}+\frac{(k+1)(2n-k)}{2}\cdot L_1.\end{aligned} \tag{19}$$

(2) 本文提出的特征选择算法的计算量分析

本文提出的特征选择算法每次入选的特征有多个，整个过程的计算量由 3 部分组成，即互信息的计算量、参数估计的计算量及 Kullback-Leibler 距离的计算量. 总的计算量为

$$C_2=O(m*n)+(IIS_{a_1}+D_{a_1})+(IIS_{a_2}+D_{a_2})+\ldots+(IIS_{a_i}+D_{a_i}). \tag{20}$$

其中 $O(m*n)$ 为互信息的计算量，a_i 为第 i 次选中的特征个数，IIS_{a_i} 和 D_{a_i} 分别表示在第 i 次特征选择时参数估计的计算量和 Kullback-Leibler 距离的计算量.

设 $IIS_{\max}$ 为在 i 次特征选择中参数估计时间的最大量，则

$$C_2\leqslant O(m*n)+(IIS_{\max}+D_{a_1})+(IIS_{\max}+D_{a_2})+\ldots+(IIS_{\max}+D_{a_i}).$$

由公式(13)可知，$D_{a_1}=D_{a_2}=\ldots=D_{a_i}$，则

$$C_2\leqslant O(m*n)+i\cdot IIS_{\max}+i\cdot D_{a_1}. \tag{21}$$

由算法可知，$m\ll n$，$i<k$ 及 $k\ll n$，$IIS_{\min}$ 与 $IIS_{\max}$ 的计算复杂度属于同一数量级，而对数似然的计算复杂度

与 Kullback-Leibler 距离的计算量大致相同，所以，本文提出的特征选择算法所需的运算量小于 Della Pietra 等人提出的特征选择算法所需的运算量.

4 改进最大熵模型的应用及实验结果

4.1 基于最大熵原理的义类排歧

作者将上面描述的建立最大熵模型的方法应用于解决汉语文本中的义类排歧问题.

模型输入：已知多义词 w 的由 N 个样本组成的样本空间：$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$，$(x_i,y_i)$表示当上下文信息为 x_i 时，w 的义类为 y_i. x_i 为 y_i 的上下文环境. 模型的目标是利用最大熵原理建立学习模型 $p(y|x)$，其含义为在上下文为 x 时输出义类为 y 的概率.

模型输出：特征集及对应参数集，即$\langle S,\lambda\rangle$；其中 $S=\{f_1,f_2,\ldots,f_n\}$且 $\lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_n\}$.

4.2 实验过程

(1) 语料库和义类词典

在实验中，将 2 000 万字已经进行了词切分和词性标注的《人民日报》语料库作为系统的数据来源，以在语料库中常出现的词性为动词或名词的高频多义词作为义类排歧的对象. 在对多义词进行语义标注时，采用的是《同义词词林》中的义类代码，代码由大、中、小三级组成，如“建”有两个义类“Hc05”和“Hd01”，它们分别代表两种不同的意义.

(2) 样本数据

根据最大熵模型，义类排歧模型的样本空间为从语料库中抽出包含某个多义词的词及其周围的上下文环境，然后对每个样本进行人工排歧，形成样本空间. 在样本数据的准备过程中，我们做了两方面的工作. (1)确定义类排歧的对象为同一词性内的义类歧义词；(2)对多义词周围所取的上下文的长度的原则定为：以句子为单位，在一句中选该词周围前后各 7 个词. 以“打”、“建”和“获”这 3 个具有义类歧义的词作为实验对象，它们在语料库中出现的次数及候选特征的个数见表 1.

(3) 特征选择和参数估计

按本文提出的特征选择算法最终产生的特征个数和 T 的值见表 2. 该模型产生的有关“打”的部分特征见表 3.

表 1

义类歧义词	样本个数	候选特征个数
打	1 642	1 145
建	1 928	3 029
获	2 682	3 766

表 2

义类歧义词	最后 T 的值	特征个数
打	1.117	377
建	1.216	402
获	1.343	285

表 3

$y=$“$Fa10$”且“井”$\in x$
$y=$“$Fa10$”且“深”$\in x$
$y=$“$Fa10$”且“扩孔”$\in x$
$y=$“$Hi44$”且“死”$\in x$
$y=$“$Hi44$”且“士兵”$\in x$
$y=$“$Hi44$”且“致残”$\in x$

(4) 用最大熵模型进行义类排歧的过程及结果

具体过程为：

- 找出含有指定多义词 w 的上下文(x,w)，其中 x 为多义词 w 的上下文环境；
- 根据模型学习到的关于 w 的参数集(S,λ)，用公式(9)和(10)计算 w 的各个义类在 x 下的条件概率 $p(y_i|x)$，其中 $y_i\in Y$.

· 取条件概率较大者对应的义类为所选义类.

在此,我们分别采用封闭测试和开放测试两种方法对模型进行测试,测试正确率定义为

$$\text{义类排歧正确率}=\frac{\text{标对义类的样本个数}}{\text{测试集中的样本个数}}.$$

得到的测试结果见表 4. 从表中数据可以看出,用本文提出的特征选择算法建立的最大熵模型可以保证有较高的排歧准确率.

表 4

	封闭测试		开放测试	
	样本个数	正确率(%)	样本个数	正确率(%)
打	100	89.5	50	83.7
建	100	93.2	50	90.6
获	100	91.8	50	89.1

5 结束语

本文提出一种改进的最大熵方法,该方法利用互信息的概念,使用 Z-测试方法进行特征选择,并以建立模型与经验模型的 Kullback-Leibler 距离作为过程的结束条件,因此,可以保证模型的准确性. 将模型用于汉语的义类排歧中,取得了较高的排歧正确率. 本文提出的方法还可用于词性标注、句子边界识别等问题.

致谢　本文的研究得到国家自然科学基金资助,此项目编号为 69433010.

参考文献

1 Ronnald Rosenfeld. A maximum entropy to adaptive statistical language learning. Computer Speech and Language, 1996, 10(3):187～228

2 Andrei Mikheev *et al*. Collocation Lattices and maximum entropy models. In: Zhou Joe ed. Proceedings of the 5th Workshop on Very Large Corpora. Beijing: Association for Computational Lingnistics, 1997. 216～230

3 Berger A L, Della Pietra S *et al*. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):40～72

4 Della Pietra S, Della Pietra V *et al*. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligent, 1997,19(4),380～393

5 Church K, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16(1),22～29

6 Frank Smadja. Retrieving collocation from text: Xtract. Computational Linguistics, 1993,19(1):143～175

An Improved Maximum Entropy Language Model and Its Application

LI Juan-zi　HUANG Chang-ning

(*Department of Computer Science and Technology　Tsinghua University　Beijing*　100084)

(*State Key Laboratory of Intelligent Technology and Systems　Tsinghua University　Beijing*　100084)

Abstract　The maximum entropy approach is proved to be expressive and effective for the statistics language modeling, but it suffers from the computational expensiveness of the model building. An improved maximum entropy approach which makes use of mutual information of information theory to select features based on Z-test is proposed. The approach is applied to Chinese word sense disambiguation. The experiments show that it has higher efficiency and precision.

Key words　Language model, maximum entropy model, parameter estimation, feature selection, mutual information, Z-test.