语言信息处理技术中的最大熵模型方法

Method of Maximum Entropy Model for Language Processing

李素建* 刘群 张志勇 程学旗
中国科学院计算技术研究所 （北京 100080）

摘 要 最大熵模型作为一种统计方法被有效地应用，可以控制细微特征，具有可重用性，简单易于理解等优点，在进行汉语信息处理的工作时我们首次引入了该模型。本文通过一个实例引入最大熵的概念，并对该框架模型进行形式化定义和描述，同时介绍了该模型中进行参数估计和特征选取的算法。

关键词 自然语言处理 最大熵模型 GIS 算法 统计方法

Abstract As a statistical method, the framework of maximum entropy is efficiently used. In its applications the accuracy is at or near the state-of-the-art. The model is easy to understand, and at the same time it can control subtle features and have reusability. The goal of this paper is to provide a brief description of formalism for the principle of the maximum entropy. And some important algorithms for parameter estimation and feature induction.are also introduced.

1 引言

进行汉语处理时经常遇到的问题有：分词、词性标注、语法和语义分析等等。这些自然语言中的问题都可以形式化为分类问题，估计某一类 y 在上下文 x 中发生的概率，即 p(y,x)。在汉语中上下文 x 的内容可以包括汉字、词、词性等，对于不同的任务上下文的内容也不同。这类问题可以采用统计建模的方法去处理。首先是采集大量样本进行训练，样本代表了该任务的知识和信息，选取样本的好坏确定了知识完整性的程度。然后建立一个统计模型，并把样本知识结合到模型中，来预测随机过程将来的行为。

在自然语言处理中有不少统计建模的例子，目前在对英语的处理中，由于最大熵模型的简洁、通用和易于移植，在统计方法中经常采用该技术[1]。汉语中词性标注和短语边界识别多使用 HMM 的统计模型[2,3]，还未见有论文或资料谈到使用最大熵的方法。本文结合汉语语言的特点，对最大熵框架模型及其使用进行介绍。第 2 节通过一个英汉翻译的简单实例引入最大熵模型，第 3 节对最大熵模型进行了形式化描述，同时介绍了估计模型参数的有效算法。第 4 节对如何为所处理问题选择特征提供了算法。第 5 节总结了该模型的应用及优点。

2 最大熵模型的简单实例

我们以英汉翻译为例：对于英语中的“take”，它对应汉语的翻译有：

(t1)“抓住”：The mother takes her child by the hand.母亲抓住孩子的手。

(t2)“拿走”：Take the book home. 把书拿回家。

(t3)“乘坐”：to take a bus to work. 乘坐公共汽车上班。

(t4)“量”：Take your temperature. 量一量你的体温。

(t5)“装”：The suitcase wouldn’t take another thing. 这个衣箱不能装别的东西了。

(t6)“花费”：It takes a lot of money to buy a house. 买一所房子要花一大笔钱。

(t7)“理解、领会”：How do you take this package? 你怎么理解这段话？

假设对于所有的英文”take”，只有这七种翻译。则存在着如下限制：

$$p(t1|x)+p(t2|x)+p(t3|x)+\ldots+p(t7|x)=1 \quad (1)$$

p(ti|x)(1≤i≤7)表示在一个含有单词 take 的英文句子中，take 翻译成 ti 的概率。

在这个限制下，对每种翻译赋予均等一致的几率为：p(t1|x)=p(t2|x)=…=p(t7|x)=1/7

但是对于“take”，我们通过统计发现它的前两种翻译(t1)和(t2)是常见的，假设满足如下条件

---

* 作者简介：李素建，博士生，研究方向为自然语言理解、机器翻译、知识挖掘

p(t1|x)+p(t2|x) = 2/5 (2)

在(1)和(2)共同限制下，分配给每个翻译的概率分布形式有很多。但是最一致的分布为：

p(t1|x)=p(t2|x)=1/5

p(t3|x)=p(t4|x)=p(t5|x)=p(t6|x)=p(t7|x)=3/25

可以验证，最一致的分布具有最大的熵值。

但是上面的限制，都没有考虑上下文的环境，翻译效果不好。因此我们引入特征。例如，英文"take"翻译为"乘坐"的概率很小，但是当"take"后面跟一个交通工具的名词"bus"时，它翻译成"乘坐"的概率就变得非常大。为了表示 take 跟有"bus"时翻译成"乘坐"的事件，我们引入二值函数：

$$f(x,y)=\begin{cases}1 & if\ \mathrm{y}="乘坐" \text{ and } \hat{}\,\mathrm{next(x)}="\mathrm{bus}" \\ 0\end{cases} \tag{3}$$

x 表示上下文环境，这里看以看作是含有单词 take 的一个英文短语，而 y 代表输出，对应着"take"的中文翻译。^next(x)看作是上下文环境 x 的一个函数，表示 x 中跟在单词 take 后的一个单词为"bus"。这样一个函数我们称作一个特征函数，或者简称一个特征。引入诸如公式(3)中的特征，它们对概率分布模型加以限制，求在限制条件下具有最一致分布的模型，该模型熵值最大。

3 最大熵模型框架

如何利用最大熵框架模型得出在特征限制下最优的概率分布，例如上面的概率值 $p(t_i|x)$。下面对该模型进行形式化描述，并介绍有关算法。

假设对于训练数据有一个样本集合为$\{(x_1,y_2), (x_1,y_2), \ldots, (x_N,y_N)\}$，每一个 $x_i(1\leqslant i\leqslant N)$表示一个上下文，$y_i(1\leqslant i\leqslant N)$表示对应的结果。对于这个训练样本，我们得到(x,y)的经验分布，定义如下：

$$\tilde{p}(x,y)=\frac{1}{N}\times \text{number of times that (x, y) occurs in the sample} \tag{4}$$

要对上面大小为 N 的训练样本集合建立统计模型，可利用的是样本集合的统计数据。模型中特征函数的引入，使模型依赖于上下文的信息。假设我们给出 n 个特征函数 f_i，对每个特征进行条件限制：期望概率值等于经验概率值，如下：

$$p(f_i)=\tilde{p}(f_i) \quad i\in\{1,2,\ldots,n\} \tag{5}$$

其中，期望值和经验值分别为：

$$p(f)\equiv\sum_{x,y}\tilde{p}(x)p(y\mid x)f(x,y) \tag{6}$$

$$\tilde{p}(f)\equiv\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{7}$$

要求得最优的 p(y|x)值，我们要得到一个最为一致(uniform)分布的模型，条件熵作为衡量一致 (uniform)的标准，

$$H(p)\equiv-\sum_{x,y}\tilde{p}(x)p(y\mid x)\log p(y\mid x) \tag{8}$$

求在限制条件下具有最大熵值的模型，C 表示所有可能满足限制条件的概率分布模型的集合，

$$p^*=\arg\max_{p\in C}H(p) \tag{9}$$

$$C\equiv\{p\in P\mid p(f_i)=\tilde{p}(f_i) \quad for\ i\in\{1,2,\cdots,n\}\} \tag{9`}$$

由 H(p)满足以下限制：

$$(1)\ p(y|x) \geq 0 \ for\ all\ x,\ y$$

$$(2)\ \sum_{y} p(y|x) = 1 \ for\ all\ x$$

$$(3)\ \sum_{x,y} \tilde{p}(x)p(y|x)f(x,y) = \sum_{x,y} \tilde{p}(x,y)f(x,y)\ for\ i \in \{1,2,...,n\}$$

求限制条件下 H(p)的最大值，为每一个特征 f_i 引入一个参数 λ_i，因此定义拉格朗日函数 ξ(p, Λ, γ)如下：

$$\xi(p,\Lambda,\gamma) \equiv -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x)$$
$$+\sum_{i} \lambda_i (\sum_{x,y} p(x,y)\tilde{f}_i(x,y) - \tilde{p}(x)p(y|x)f_i(x,y))$$
$$+\gamma(\sum_{x} p(y|x) - 1)$$

参数 γ 和 Λ={λ_1, λ_2, … , λ_n}对应着 n+1 个限制。

保持 γ 和 Λ 不变，计算拉格朗日函数 ξ(p, Λ, γ)在不受限情况下的最大值，就可以得到 p(y|x)的最优值，因此对 p(y|x)求导得到：

$$\frac{\partial \xi}{\partial p(p|x)} = -\tilde{p}(x)(1+\log p(y|x)) - \sum_{i} \lambda_i \tilde{p}(x) f_i(x,y) + \gamma \qquad (10)$$

满足公式(10)为 0 的参数 $p^*(y|x)$最优，使函数 ξ(p, Λ, γ)得到最大值，并使 H(p)得到限制条件下的最大值，此时，$p^*(y|x)$为：

$$p^*(y|x) = \exp(\sum_{i} \lambda_i f_i(x,y))\exp(-\frac{\gamma}{\tilde{p}(x)} - 1) = Z(x)\exp(\sum_{i} \lambda_i f_i(x,y))$$

根据对于所有的 x，满足 $\sum_{y} p_\lambda(y|x) = 1$，所以范化因子 Z(x)为

$$Z_\lambda(x) = \sum_{y} \exp(\sum_{i} \lambda_i f_i(x,y)) \qquad (11)$$

这样 γ^*和 p^*就转化成 Λ^*的函数，要求最优值 Λ^*，令

$$\Psi(\Lambda) = \xi(p^*, \Lambda, \gamma^*) \qquad (12)$$

$$\lambda^* = \arg\max_{\lambda} \Psi(\lambda) \qquad (13)$$

在参数估计中求使 Ψ(λ)最大的最优值 Λ^*，不可能分析得到。由于 Ψ(λ)对于参数 λ 是向上∩型的函数，一般采用数值计算方法。在最大熵模型中常采用的优化方法是[Darroch and Ratcliff, 1972]的迭代缩放方法(iterative scaling)，即 Generalized Iterative Scaling (GIS)，用来得到具有最大熵分布的参数值 Λ^*。

GIS 算法如下：(N 表示样本总数，n 表示特征总数)

1. 初始化参数 $\lambda_i^{(1)}$ 为任意值(例如 0)，计算 $d_i = E_{\tilde{p}}(f_i(y,x)) = \Sigma_{y,x} \tilde{p}(x,y)f_i(y,x) = 1/|N| \Sigma_{t=1..N} f_i(y_t,x_t)$，i=1..n+1；
2. 设置迭代次数 k=1；
3. 计算当前模型分布的期望值

   对于所有的期望限制：

$$E_{p(n)}(f_i) = \Sigma_{y,x} p(x,y)f_i(y,x) \cong 1/|N| \Sigma_{t=1..N} \Sigma_Y p^{(n)}(y|x_t)f_i(y,x_t) \qquad (\text{based on } p^{(n)}(y|x_t))$$

$$p^{(n)}(y|x) = (1/Z)\, e^{\Sigma_{i=1..n} \lambda_i f_i(y,x)}$$

4. 修改 $\lambda_i^{(n+1)} = \lambda_i^{(n)} + (1/c)\log(d_i/e_{p(n)}(f_i))$，$c = \forall y,x \Sigma_{i=1..N} f_i(y,x)$, k=k+1

5. 反复执行 3、4，直到收敛或迭代次数 k 大于某一个数 100 [4]

对于参数值的计算，Darroch[5]证明 GIS 算法参数是收敛的。在文献[6，7]中又介绍了一个改进的迭代缩放方法，该算法在计算模型参数时加快了收敛速度。

4 特征引入算法

我们在最大熵模型中要选取特征，尤其对于自然语言问题，这些特征的选取往往具有很大的主观性。同时因为特征数量之大，我们必须引入一个客观标准自动计算以引入特征到模型中，同时针对特征进行参数估计。Della Pietra et al[6] 对自然语言处理中随机域(random field)的特征选取进行了描述，这里进行特征选取时，是由特征的信息增益作为衡量标准的。一个特征对所处理问题带来的信息越多，该特征越适合引入到模型中。一般我们首先形式化一个特征空间，所有可能的特征都为候补特征，然后从这个候补特征集合内选取对模型最为有用的特征集合。特征引入的算法如下：

算法 2: 特征引入(Feature Induction，简称 FI)算法

输入：候补特征集合 F，经验分布 $\tilde{p}(x,y)$

输出：模型选用的特征集合 S，结合这些特征的模型 P_S

1. 初始化：特征集合 S 为空，它所对应的模型 P_s 均匀分布，n=0；
2. 对于候补特征集合 F 中的每一个特征 f∈F，计算加入该特征后为模型带来的增益值 G_f;
3. 选择具有最大增益值 G(S,f)的特征 f_n;
4. 把特征 f_n 加入到集合 S 中，$S=(f_1, f_2, \ldots, f_n)$;
5. 重新调整参数值，使用 GIS 算法计算模型 P_s;
6. n=n+1，返回到第 2 步；

该算法的第 2 步中，要计算每个特征的增益值，这里是根据 Kullback-Leibler（简称 KL）距离来计算的。衡量两个概率分布 p 和 q 的 KL 距离，公式如下:

$$D(p \| q) = \sum_x p(x) \ln \frac{p(x)}{q(x)} \tag{14}$$

距离的大小与两种分布的相似程度成反比，距离越小表示两种分布越逼近。因此在加入第 n 个特征前后，求的模型分布与样本分布之间的 KL 距离为：

$$D(\tilde{p} \| p^{(n-1)}) = \sum_x \tilde{p}(x) \ln \frac{\tilde{p}(x)}{p^{(n-1)}(x)}$$

$$D(\tilde{p} \| p^{(n)}) = \sum_x \tilde{p}(x) \ln \frac{\tilde{p}(x)}{p^{(n)}(x)}$$

这样，我们定义引入第 n 个特征 $f^{(n)}$后的增益值为：

$$G(p, f^{(n)}) = D(\tilde{p} \| p^{(n-1)}) - D(\tilde{p} \| p^{(n)}) \tag{15}$$

因此选择的第 n 个特征为：

$$f^{(n)} = \arg\max_{f \in F} G(p, f^{(n)}) \tag{16}$$

模型选取特征是从特征空间这里看作是候补特征集合中选取的，因此我们首先定义一个特征空间，也可以看作是选取特征的模式。特征空间的定义和选取对于问题的处理也是一个关键。目前经过加工的语料有词，词性标注或语法信息都包含在语料中，我们根据不同的任务定义不同的特征空间。例如，在词性标注时我们定义特征空间，这里选取当前词 w_i 的前后两词及前两词的词性，即：{w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1},

t_{i-2} }，然后特征空间具体化，如果将所有特征都考虑在内，则数量极其庞大，这样我们就可以根据上面介绍的 FI 算法选择一个最佳的特征集合加入到最大熵模型中。最后对于选取的每个特征都赋予一个权值，也就是我们模型的参数。这时就可以根据模型进行预测将来的行为。由于篇幅有限，本文不再详细介绍。

5 结束语

应用最大熵模型的统计方法可以把模型与语言学知识分为两个模块进行处理，在机器翻译[8]、词性标注[9]、语法分析[1]、部分分析[10]等各项语言处理任务中都得到了有效的应用，模型可以被多次利用，对不同的任务只是选择的特征不同。最大熵框架的这种通用性和重用性允许实验者使用其他任务中相同的参数估计程序。参数估计的代码实质上与特定任务无关，一个实现就可以满足所有的其他任务模型。更重要的是，最大熵模型在每个任务中都表现的相当好，尽管所有任务在本质和复杂性上是不同的。各种实验结果表明研究者可以使用和重用最大熵框架到非常广泛的任务中去，并具有很高的准确性。此外，最大熵模型本身对于任务的处理还有以下优点：

1) 控制细微结果。使用最大熵可以准确为变量间的细微依赖关系建模，在建立最大熵模型时，可以跨距离地选取特征。这样选取的特征用传统的预测模型技术是不可能的。
2) 不做未经验证的假设。最大熵承认已有的事实，对所选特征没有独立性假设。而传统的预测模型例如决策树、逻辑回归和神经网络对信息都会做一些错误的假设，
3) 该模型简单、易于理解。本文中的特征集不需要深层的语言学知识，只问上下文中的基本问题。比其他方法较少依赖语言学知识、预处理、或语义数据库，因此更容易指定和移植特征。尽管特征明显的简化，它们仍可以有效地近似复杂的语言学关系。

本文对最大熵模型及其算法进行了详细的介绍，对于汉语处理中的各项任务，需具体问题具体分析，分别选取不同的特征集合结合到该模型中，而无须过多涉及到模型和算法问题，因此该模型框架对于汉语信息处理具有实际的意义。

参考文献

[1] Skut, Wojciech and Thorsten Brants, A Maximum Entropy Partial Parser for Unrestricted Text, In 6th Workshop on Very Large Corpora, Montreal, Canada, August, 1998.

[2] 周强, 规则和统计相结合的汉语词类标注方法, 《中文信息学报》 , 9（3）,1995

[3] 周强, 一个汉语短语自动界定模型，软件学报第7卷，增刊，315-322, 1996

[4] Adwait Ratnaparkhi, Maximum Entropy Models for Natural Language Ambiguity Resolution, Ph.D. Dissertation, University of Pennsylvania, 1998.

[5] Darroch, J.N. and Ratcliff, D., Generalized Iterative Scaling for Log-Linear models, Annals of Mathematical Statistics, 43(5): 1470-1480, 1972.

[6] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty, Inducing features of random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 19:4, pp.380--393, April, 1997.

[7] Adam Berger, The Improved Iterative Scaling Algorithm: A Gentle Introduction, http://www.cs.cmu.edu/afs/cs/user/aberger/www/ps/scaling.ps, 1997.

[8] Adam Berger, Stephen Della Pietra, and Vincent Della Pietra, A maximum entropy approach to natural language processing, Computational Linguistics, (22-1), March 1996.

[9] Ratnaparkhi, A., A maximum entropy model for part-of-speech tagging. In Proceeding of the Conference on Empirical Methods in Natural Language Processing, 1996.

[10] Erik F. Tjong Kim Sang and Sabine Buchholz, Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000.