应用粒子群优化-条件随机域的文本生物实体识别

豆增发，高琳
（西安电子科技大学计算机学院，710071，西安）

摘要：针对生物医学文本中传统生物实体识别算法的精确度不高的问题，提出了一种新的基于粒子群优化-条件随机域的生物实体识别算法.新算法利用改进的粒子群优化算法训练条件随机域模型，并将训练后的条件随机域模型应用到生物实体的识别上.改进的粒子群优化算法引入粒子群聚集度来防止粒子群过早地陷入局部收敛，用迭代间对数似然相对变化率来控制算法的收敛，用线性变化的惯性因子和学习因子来控制搜索范围.实验结果表明，基于改进粒子群优化的条件随机域模型较隐马尔科夫模型、最大熵马尔科夫模型、支持向量机以及传统条件随机域模型等方法具有更高的精确率和召回率.

关键词：条件随机域模型；粒子群优化；粒子群聚集度；对数似然相对变化率；生物实体识别
中图分类号：TP301 **文献标志码**：A **文章编号**：0253-987X(2010)12-0038-06

A Bio-Entity Recognition Algorithm for Literature by Conditional Random Field Model Based on Improved Particle Swarm Optimizer

DOU Zengfa, GAO Lin
(School of Computer Science, Xidian University, Xi′an 710071, China)

Abstract: A new bio-entity recognition algorithm is proposed to improve the precision of bio-entity recognition for biomedical literature. The new algorithm trains the conditional random field model using an improved particle swarm optimizer, and then applies the trained conditional random field model to bio-entity recognition. The aggregation degree of particle swarm is utilized to control the early local convergence of the particle swarm optimizer, the relative change ratio of log-likelihood between iterations is employed to end its iterations, and the inertia factor and learning factor are set as linear variables to control the scope of search space. Experimental results show that the proposed algorithm outperforms the models of HMM, MEMM, SVM and traditional L-BFGS CRF on precision and recall.

Keywords: conditional random field model; particle swarm optimizer; aggregation degree of particle swarm; relative change ratio of log-likelihood; bio-entity recognition

生物实体识别的目标是从大量的生物医学文本数据中识别和分类生物学家感兴趣的名称或符号，例如，从MEDLINE数据库中自动标注出所有的基因、蛋白质名称.近年来，条件随机域模型[1]在生物实体识别等方面获得了成功的运用[2-5]，该模型放宽了隐马尔科夫模型的强随机假设，有效地克服了标注偏置的问题.

在条件随机域模型中，最大似然参数的估计关系到条件随机域模型的性能和精确度[1].本文利用改进粒子群优化算法来估计条件随机域模型的最大似然参数.为了防止粒子群优化在搜索早期陷入局部收敛，使用一个合适的粒子群聚集度对收敛进行

收稿日期：2010-06-08. 作者简介：豆增发(1979—)，男，博士生，系统分析师；高琳(联系人)，女，教授，博士生导师.
基金项目：国家自然科学基金重点资助项目(60933009)；高等学校博士学科点专项科研基金资助项目(200807010013)；国家自然科学基金资助项目(60970065).

控制. 为了克服粒子群优化在最佳位置附近可能陷入无穷迭代的问题,采用迭代间的对数似然相对变化率来终止迭代. 不同于传统算法,对于粒子群优化的惯性因子和学习因子,采用了线性变化值,将训练后的条件随机域模型在 GENIA 资料库[6]、GENETAG 资料库[7]以及自建资料库上进行了评价,结果表明,经改进粒子群优化算法训练后的条件随机域模型较隐马尔科夫模型、最大熵马尔科夫模型、支持向量机以及传统条件随机域模型具有更高的精确率和召回率.

1　条件随机域模型

条件随机域模型是用于计算在给定输入序列的前提下输出序列的条件概率的一种无向图模型. Lafferty [1]给出了条件随机域模型的定义. 在给定某一观测序列 x 的前提下,某一特定标签序列 y 的概率为位函数的归一化乘积,如下式所示

$$p(y \mid x,\boldsymbol{\lambda}) = \frac{1}{Z(x)}\exp\Big(\sum_j \lambda_j F_j(y,x)\Big) \quad (1)$$

式中:$F_j(y,x)$可作为状态函数 $s(y_{i-1},y_i,x,i)$或过渡函数 $t_j(y_{i-1},y_i,x,i)$;λ_j 可作为一个权重来表示特征 f_j 的精确度;$\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots)$;$Z(X)$是归一化因子

$$Z(x) = \sum_{x,y}\exp\Big(\sum\sum_k \lambda_k f_k(c,y_c,x)\Big) \quad (2)$$

式(1)那样的条件随机域模型的最大似然参数估计问题,就是通过一个训练数据集合 $D=\{(y^{(1)},x^{(1)}),\cdots,(y^{(n)},x^{(n)})\}$使条件随机域模型的对数似然函数最大化,从而得出一组合适的权重参数 λ. 条件随机域模型的对数似然函数定义为[1]

$$L(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\log(p(y_i \mid x_i)) - \sum_{j=1}^{n}\frac{\lambda_j^2}{2\sigma^2} \quad (3)$$

从数值优化的角度来看,条件随机域的对数似然函数在整个参数空间是平滑和凹形的,即如果 $\boldsymbol{\lambda}$ 的值使得对数似然函数取得全局极大值,则对数似然函数对 $\boldsymbol{\lambda}$ 的偏导为 0.

2　基于改进粒子群优化的条件随机域模型参数估计方法

2.1　改进的粒子群优化算法

粒子群优化算法[8-9]通过初始化一个粒子群(x_i)并使其按一定规律更新进化以搜索最优值. 每个粒子的速率和位置根据下式更新

$$v_{id}(t+1) = w\,v_{id}(t) + c_1\,\mathrm{rand}()\,(p_{id}-x_{id}(t)) + c_2\,\mathrm{rand}()\,(p_g - x_{id}(t)) \quad (4)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (5)$$

式中:$v_{id}(t)$表示粒子 i 在维度 d 上的速率;$x_{id}(t)$表示粒子 i 在维度 d 上的当前位置;$x_{id}(t+1)$表示粒子 i 在维度 d 上的下一个位置;w 是惯性因子;c_1 是局部学习因子;c_2 是全局学习因子;p_{id} 表示粒子 i 在维度 d 上迄今所得到的最优值;p_g 表示所有粒子在整个搜索空间中迄今所得到的最优值;rand()是值域为[0,1]的随机函数.

为了防止搜索中过早地出现局部收敛的情况,采用粒子群聚集度[10]来进行控制. 粒子群聚集度描述粒子群的分散情况,定义为[10]

$$d(t) = \max\{\,|\,x_{id}-x_{jd}\,|,\quad i,j=1,2,\cdots,m;\ i\neq j;d=1,2,\cdots,N\} \quad (6)$$

式中:m 表示粒子群大小;N 表示搜索空间的维度;x_{id}表示维度 d 上的第 i 个粒子;x_{jd}表示维度 d 上的第 j 个粒子.

为了避免在最佳位置附近陷入无限迭代,采用迭代间对数似然相对变化率[11]来决定是否终止迭代. 粒子群的迭代间对数似然相对变化率定义为[11]

$$L(x_{id}) = \frac{\lg(x_{id}(t+1)) - \lg(x_{id}(t))}{\lg(x_{id}(t+1))} \quad (7)$$

式中:当 $L(x_{id})$小于门限值 10^{-7} 或者其他定义值时,强制终止迭代.

惯性因子 w 决定算法的搜索范围,一般认为较大的惯性因子对全局搜索有利,而较小的惯性因子对局部搜索有利. 因此,我们使惯性因子 w 根据式(8)线性递减,使算法在搜索初期较好的全局搜索能力,而在搜索末期具有具有较好的局部搜索能力.

$$w(t) = w_{\mathrm{start}} - \frac{w_{\mathrm{start}} - w_{\mathrm{end}}}{M}\,t \quad (8)$$

式中:M 表示迭代的最大次数;t 表示当前的迭代次数. 为了使算法在开始阶段具有较强的全局学习能力,而在末期具有较强的局部学习能力,起初设 c_1 和 c_2 都为 2,然后它们根据式(9)线性变化.

$$c_1 = 2+\frac{t}{M};\quad c_2 = 2-\frac{t}{M}\quad (t=0\sim M-1) \quad (9)$$

2.2　基于改进粒子群优化的条件随机域模型参数估计方法

在训练条件随机域模型时,将条件随机域模型的对数似然函数作为适应函数,将参数向量 $\boldsymbol{\lambda}$ 作为

粒子群，并使它们在一个 d 维空间搜索最优值. 详细步骤如下.

步骤 1 设当前的迭代次数 $t=0$，$v_{id}(0)=0$，搜索空间维数为 d. 初始化包含 m 个粒子的种群 X

$$\boldsymbol{x}_{id}=\{\lambda_{1d},\lambda_{2d},\cdots,\lambda_{nd}\}=\{r_1,r_2,\cdots,r_n\}$$
$$(i=0\sim m-1)$$

设局部最优位置和全局最优位置为 $p_{id}=x_{id}$，$p_g=0$. 将局部最佳适应值 f_{id} 保存在数组 $F[i][d]$，将全局最佳适应值保存在变量 F_g. 惯性因子的一个经验值是搜索初期取值接近 1 而在搜索末期取值接近 0.4[9]，因此，设 $w_{\text{start}}=0.9$，$w_{\text{end}}=0.4$.

步骤 2 计算每个粒子的适应值

$$f_{id}=\sum_{i=1}^{n}\log(p(y_i\mid x_i))-\sum_{j=1}^{n}\frac{\lambda_j^2}{2\sigma^2}$$

步骤 3 与数组 $F[i][d]$ 比较对应的每个粒子的适应值. 如果当前值大于 $F[i][d]$，则当前值为局部最优值，并将 $F[i][d]$ 重置为当前值. 此外，如果当前值大于 F_g，则设当前值为全局最优值并重置 F_g 为当前值.

步骤 4 更新粒子群的搜索速率和位置

$$v_{id}(t+1)=\left(w_{\text{start}}-\frac{w_{\text{start}}-w_{\text{end}}}{M}t\right)v_{id}(t)+$$
$$\left(2+\frac{t}{M}\right)\text{rand}()\ (p_{id}-x_{id})+$$
$$\left(2-\frac{t}{M}\right)\text{rand}()\ (p_g-x_{id})$$

如果下一个速率 $v_{id}(t+1)$ 大于速率阈值 $V_{\max}$，则设 $v_{id}(t+1)=V_{\max}$.

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1)$$

步骤 5 如果 t 对 m 取模 $t\%m=0$，计算粒子群的聚集度 $d(t)$

$$d(t)=$$
$$\max\{|x_{id}-x_{jd}|,i,j=1,2,\cdots,m;i\neq j\}=$$
$$\max\{[(\lambda_{di1}-\lambda_{dj1})^2+\cdots+(\lambda_{din}-\lambda_{djn})^2]^{1/2},$$
$$i,j=1,2,\cdots,m;i\neq j\}=$$
$$\max\{[\sum_{k=1}^{n}(\lambda_{dik}-\lambda_{djk})^2]^{1/2},$$
$$i,j=1,2,\cdots,m;i\neq j\}$$

如果聚集度 $d(t)$ 小于设定的门限值 e，则在 d 维空间上重新初始化粒子群的速率和位置.

步骤 6 $t=t+1$.

如果 $L(x_{id})<10^{-7}$ 或者 $t=M-1$，则终止算法，否则转向步骤 2.

3 实验结果

我们在 GENIA 资料库、GENETAG 资料库以及自建资料库上进行了生物实体识别实验，用精确率 P、召回率 R 以及 F 值对算法性能进行评价. 为了验证 PSO 算法的粒子群大小和维度是否合适，设粒子群的大小为 20、40、80 和 160，对每一组种群，分别使用 10、20、303 种不同的维度来测试. 为了获得较为合理的似然参数，对每一组维度-粒子群数运行 50 次，然后取参数的平均值，以抵消进化算法本身的随机性. 实验结果表明，基于改进粒子群优化的条件随机域模型在粒子群规模较小时，它的性能劣于隐马尔科夫模型，而当粒子群规模较大，如超过 40 时，它的性能总体上优于隐马尔科夫模型、最大熵马尔科夫模型、支持向量机以及传统的条件随机域模型. 当粒子群的大小是 160、维度是 30 时，基于改进粒子群优化的条件随机域模型的性能为最佳. 通用的精确率 P、召回率 R 和 F 值的定义如下

$$P=\frac{m_{\text{c}}}{m_{\text{all}}}\tag{10}$$

$$R=\frac{m_{\text{i}}}{m_{\text{all}}}\tag{11}$$

$$F-s=\frac{2PR}{P+R}\tag{12}$$

式中：m_{c} 表示正确标注的生物实体数量；m_{i} 表示可以被识别的生物实体名称数量；m_{all} 表示参与实验的总的生物实体数量.

在基于改进粒子群优化的条件随机域模型中，我们使用了 6 213 个语义特征，它们中的多数符合 Settles 使用的特征[3]. 除了语义特征，还使用了如表 1 所示的英文拼写特征. 另外，从训练数据中抽取了一些生物实体特征，比如“peri-kappa_ B_ site”、“human_ immunodeficiency_ virus_ type_ 2_ enhancer”、“monocyte”、“T_ cell”等一些常用的基因名称[12].

3.1 GENIA 资料库上的实验结果

GENIA 资料库[6]是一个从美国国家图书馆医学数据库中抽取的有标注的摘要资料库. 在 GENIA 资料库中，用 XML(GPML)标注了摘要中参与蛋白质反应的实体及其属性. GENIA 资料库 3.02 版包含了 2 000 个实体，我们从中选择 1 500 个摘要作为训练数据，剩下的 500 个摘要作为测试数据. 测试结果如表 2 所示.

表 1　实验中采用的英文拼写特征

英文拼写特征	正则表达式
Init Caps	[A−Z]. *
Init Caps Alpha	[A−Z][a−z] *
All Caps	[A−Z]+
Caps Mix	[A−Za−z]+
Has Digit	. *[0−9]. *
Single Digit	[0−9]
Double Digit	[0−9][0−9]
Natural Number	[0−9]+
Real Number	[−\+][[0−9]+[\. ,]+[0−9]. ,]+
Alpha-Numeric	[A−Za−z0−9]+
Roman	[ivxdlcm]+|[IVXDLCM]+
Has Dash	. * −. *
Init Dash	−. *
End Dash	. * −
Punctuation	[,\. ;:\?! −\+"]
Greek	(alpha|beta|…|omega)
Has Greek	. *\b(alpha|beta|…|omega)\b. *
Mutation Pattern	\w * \d+− * \D+

表 2　GENIA 库上不同算法的性能比较

参数估计方法	大小	维度	P	R	F 值
隐马尔科夫模型（HMM）			0.827	0.770	0.797
最大熵马尔科夫模型（MEMM）			0.813	0.724	0.766
支持向量机（SVM）			0.793	0.722	0.756
条件随机域模型（L-BFGS CRF）			0.828	0.767	0.796
基于改进粒子群优化的条件随机域模型（PSO-CRF）	20	10	0.825	0.760	0.791
		20	0.826	0.766	0.794
		30	0.820	0.766	0.795
	40	10	0.828	0.770	0.798
		20	0.829	0.769	0.798
		30	0.830	0.769	0.798
	80	10	0.833	0.771	0.800
		20	0.832	0.772	0.801
		30	0.836	0.773	0.803
	160	10	0.839	0.774	0.805
		20	0.840	0.777	0.807
		30	0.843	0.780	0.813

3.2　GENETAG 资料库上的实验结果

GENETAG 资料库[7]是由 Lorraine Tanabe 等建立的一个基因和蛋白质资料库，用于文本中基因和蛋白质名称标注的训练. 该库的数据来自 MEDLINE 的文章摘要. 该资料库由 20 000 个随机选取的句子组成，每个句子的基因和蛋白质名称用一定的规则做了标注，然后进行了手工调整. GENETAG 资料库的最新版本是 GENETAG-05. 我们利用 GENETAG 库的训练数据对条件随机域模型做了训练，用 GENETAG 库的测试数据做了测试，并与其他算法做了比较，结果如表 3 所示.

表 3　GENETAG 库上不同算法的性能比较

参数估计方法	大小	维度	P	R	F 值
隐马尔科夫模型（HMM）			0.822	0.780	0.800
最大熵马尔科夫模型（MEMM）			0.810	0.723	0.764
支持向量机（SVM）			0.781	0.724	0.751
条件随机域模型（L-BFGS CRF）			0.813	0.766	0.789
基于改进粒子群优化的条件随机域模型（PSO-CRF）	20	10	0.824	0.770	0.796
		20	0.823	0.771	0.796
		30	0.824	0.772	0.797
	40	10	0.833	0.773	0.802
		20	0.828	0.769	0.797
		30	0.830	0.772	0.780
	80	10	0.834	0.774	0.802
		20	0.835	0.776	0.804
		30	0.836	0.778	0.806
	160	10	0.840	0.780	0.809
		20	0.842	0.780	0.810
		30	0.849	0.787	0.817

3.3　自建资料库上的实验结果

我们从 PubMed 随机选取了 200 个摘要，手工标注出疾病、基因、蛋白质等生物实体名称，构建了一个自建资料库. 用自建资料库中的 100 个摘要作为训练数据，另外 100 个摘要作为测试数据，对隐马尔科夫模型、最大熵马尔科夫模型、支持向量机、传统条件随机域模型以及基于改进粒子群优化的条件

随机域模型进行了比较,结果如表 4 所示.

表 4 自建库上不同算法的性能比较

参数估计方法	大小	维度	P	R	F 值
隐马尔科夫模型(HMM)			0.711	0.669	0.689
最大熵马尔科夫模型(MEMM)			0.703	0.614	0.655
支持向量机(SVM)			0.671	0.616	0.642
条件随机域模型(L-BFGS CRF)			0.704	0.655	0.679
基于改进粒子群优化的条件随机域模型(PSO-CRF)	20	10	0.714	0.661	0.686
		20	0.712	0.661	0.686
		30	0.714	0.663	0.688
	40	10	0.722	0.662	0.691
		20	0.719	0.659	0.689
		30	0.721	0.663	0.691
	80	10	0.724	0.664	0.693
		20	0.725	0.667	0.695
		30	0.726	0.668	0.696
	160	10	0.731	0.673	0.701
		20	0.735	0.679	0.706
		30	0.739	0.702	0.720

4 结 论

本文利用改进粒子群优化算法来估计条件随机域模型的最大似然参数.为了防止粒子群优化在早期陷入局部收敛,引入粒子群聚集度对算法进行收敛控制.为了避免粒子群优化在最佳位置陷入无限迭代,引入迭代间对数似然相对变化率作为停止标准.为了使粒子群优化获得最佳搜索效果,采用自适应变化的方式来决定惯性因子和学习因子.通过在 GENIA 资料库、GENETAG 资料库以及自建资料库上的比较实验,证明基于改进粒子群优化的条件随机域模型比隐马尔科夫模型、最大熵马尔科夫模型、支持向量机以及传统条件随机域模型具有更高的精确率和召回率.

粒子群优化算法中的群体规模和搜索空间维度一般通过实验获得.本文中,当粒子群规模为 160、维度为 30 时,获得的参数估计值为最佳.但是,在实际工程应用中,粒子群的规模和维度的取值会取决于训练数据的数量和质量,因此,实际的取值会权衡精确率和计算复杂度,如何找到这两者之间的最佳均衡点,是我们今后研究的一个方向之一.

参考文献:

[1] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th International Conference on Machine Learning. Williamstown, Massachusetts, USA: Morgan Kaufmann, 2001:282-289.

[2] KINJO AR, ROSSELLO F , VALIENTE G . Profile conditional random fields for modeling protein families with structural information [J]. Biophysics, 2009,5: 37-44.

[3] SETTLES B. Biomedical named entity recognition using conditional random fields and novel feature sets [C]// Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. New Jersey, USA: Association for Computational Linguistics, 2004:104-107.

[4] BUNDSCHUS M, DEJORI M, STETTER M, et al. Extraction of semantic biomedical relations from text using conditional random fields [J]. BMC Bioinformatics, 2008,9: 207-220.

[5] 李国臣,王瑞波,李济洪. 基于条件随机场模型的汉语功能块自动标注[J]. 计算机研究与发展, 2010,47(2):336-343.
LI Guochen, WANG Ruibo, LI Jihong. Automatic labeling of Chinese functional chunks based on conditional random fields model [J]. Journal of Computer Research and Development, 2010,47(2):336-343.

[6] KIM J D, OHTA T, TATEISI Y, et al. GENIA corpus: a semantically annotated corpus for bio-textmining [J]. Bioinformatics, 2003,19(S1): i180-i182.

[7] TANABE L, XIE N, THOM L H, et al. Genetag: a tagged corpus for gene/protein named entity recognition [J]. BMC bioinformatics, 2005,6(S1): 1-7.

[8] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of the 14th International Conference on Neural Networks. Piscataway, NJ, USA: IEEE Service Center,1995: 1942-1948.

[9] EBERHART R, KENNEDY J . A new optimizer using particle swarm theory [C]// Proceedings of the 6th International Symposium on Micro Machine and Human Science. Piscataway, NJ, USA: IEEE, 1995: 39-43.

[10] YANG Guangyou. A modified particle swarm optimizer algorithm [C]// Proceedings of the 8th International Conference on Electronic Measurement and Instruments. Piscataway, NJ, USA: IEEE, 2007: 2675-2679.

[11] MALOUF R. A comparison of algorithms for maximum entropy parameter estimation [C]// Proceedings of The 6th Conference on Natural Language Learning. New Jersey, USA: Association for Computational Linguistics, 2002: 49-55.

[12] 崔舒宁，朱丹军，冯博琴，等. 结合受控词汇表的生物基因本体标注与分类 [J]. 西安交通大学学报，2008，42(2)：171-174.
CUI Shuning, ZHU Danjun, FENG Boqin, et al. Triage and annotation of biological gene's ontology combined controlled glossary [J]. Journal of Xi'an Jiaotong University, 2008, 42(2): 171-174.

[13] 吕强，刘士荣. 一种信息充分交流的粒子群优化算法 [J]. 电子学报，2010，38(3)：664-667.
LV Qiang, LIU Shirong. A particle swarm optimization algorithm with fully communicated information [J]. Acta Electronica Sinica, 2010, 38(3): 664-667.

[14] 朱海，王宇平，权义宁，等. 采用离散粒子群算法的网格任务安全级调度 [J]. 西安交通大学学报，2010，21(6)：21-26.
ZHU Hai, WANG Yuping, QUAN Yining, et al. Security scheduling of grid tasks based on the discrete particle swarm algorithm [J]. Journal of Xi'an Jiaotong University, 2010, 21(6): 21-26.

[15] 王峰，邢科义，徐小平. 系统辨识的粒子群优化方法 [J]. 西安交通大学学报，2009，43(2)：116-120.
WANG Feng, XING Keyi, XU Xiaoping. A system identification method using particle swarm optimization [J]. Journal of Xi'an Jiaotong University, 2009, 43(2): 116-120.

（编辑　武红江）