条件随机域模型和实验分析

欧阳佑[1] 李素建[2]

(1. 北京计算语言学研究所, 北京, 100871; 2. 北京计算语言学研究所, 北京, 100871)

摘 要：条件随机域模型通过计算标注序列在观测序列下的条件概率进行标注，解决了传统模型（如隐马尔科夫模型和最大熵马尔科夫模型等）中存在的标注偏差问题，得到了更好的标注效果。时间信息识别是文本序列标注的一个典型代表，本文旨在通过 CRF 在英文时间信息识别的应用，从理论上深入分析，以及进行对比实验比较条件随机域与其他标注模型的效果，结果验证了理论和实验的一致性，CRF 更适合解决序列标注问题。

关键词：条件随机域；序列标注；时间信息识别

Conditional Random Field for Temporal Expression Recognition

OUYANG You[1] LI Sujian[2]

(1,2. Institute of Computational Linguistics,Peking University, Beijing, 100871)

Abstract: This paper introduces a probabilistic framework for sequence labeling - conditional random field (CRF). CRF carries out sequence labeling via computing a conditional probability distribution over label sequences given a observation sequence. The "label bias problem" is solved, with performance better than HMMs and MEMMs. Timex expression recognition is a typical problem of sequence labeling. We compare CRF with some other models like SVM in both theory and experiments in timex expression recognition. Comparative experiments on TIDES corpus show that CRF is more adaptive than SVM in sequence labeling problems.

Keywords: Conditional Random Field; Sequence Labeling; Temporal Expression Recognition

1 引言

条件随机域(Conditional Random Field，后面简称CRF)，是一种基于无向图的概率模型[1]，主要用于处理序列标注(sequence labeling)问题。CRF目前被广泛应用于自然语言处理任务中，如切分和词性标注[1]、组块分析[2]、实体识别[3]等，并且获得了令人满意的结果，是当前机器学习领域最热门的方法之一。

时间信息处理是文本信息挖掘的一种，其任务是获取文档中表达时间信息的词或词串，包括显式的时间表达式和各种时间前缀、后缀等等。时间信息处理分为识别和标准化两个步骤：提取出时间信息，并对其进行标准化(Normalization)得到具体描述的时间。其中时间识别是典型的序列标注任务，采用方法主要分为基于规则的方法

基金资助：本文相关研究得到 973 课题“文本内容理解的数据基础 2004CB318102”、北京市自然科学基金项目 4052019 的支持。

作者简介：欧阳佑：83年生，男，硕士生，研究方向为：信息提取，自然语言处理，email: oyangu@pku.edu.cn

李素建：75 年生，女，讲师，博士，研究方向为：计算语言学，信息提取，email: lisujian@pku.edu.cn

和基于统计的方法，目前都取得了不错的结果[8]。本文希望通过把CRF应用到时间识别这一典型任务上，以及与其他方法的结果进行比较，对CRF有更深入的认识，了解CRF与一些常见标注算法的区别和优势。

2 CRF 简介

2.1 序列模型的发展

序列数据标注领域中，最为广泛使用的是隐马尔科夫模型（后面简称HMM），HMM被成功地应用于自然语言处理，生物信息学等很多领域，取得了良好的效果。但由于HMM模型具有前提条件“独立性假设”：每个观测状态只与其对应的隐藏状态相关；而这个假设在实际问题中往往不成立，使得HMM对许多问题的处理结果并不理想。随后，将最大熵的思想应用于马尔科夫模型中，得到了最大熵马尔科夫模型(后面简称MEMM) [14]，成功地摆脱了独立性假设，相对HMM有很大的优势。但是MEMM等基于邻接状态的非生成有限状态模型，都存在被称为“标注偏差问题”的缺点，这是由于MEMM等模型中的局部概率泛化引起的。CRF则在MEMM的基础上又前进了一步，它不再单独计算每个节点标注的条件概率分布，而是计算整个标注序列在整个观测序列下的条件概率分布，解决了“标注偏差问题”。

2.2 CRF简介

CRF的主要思想是计算整个标注序列在观测序列下的条件概率，与一般的序列模型(如HMM等)不同的是，标注序列的结构可以是一般的无向图，而不仅仅限于单条有序链。由于在时间信息识别中，链图模型足以表达标注序列的结构，所以下面的公式都是针对链图模型而言的。

用随机变量$X(X_1,X_2,...,X_n)$表示待标注的数据序列，$Y(Y_1,Y_2,...,Y_n)$表示可能的标注序列，CRF需要计算的是条件概率$P(Y|X)$，然后通过这个概率选择最佳的Y。将标注序列中的各个变量Y_i根据其独立关系表示为一个无向图G(V,E)，那么全局概率$P(Y|X)$可以表示成关于G中所有的团(图的最大完全子图)的势函数(potential function)的乘积，对于链图而言，由于每个点只与其相邻点有关联，那么只有相邻的两个点的集合构成团，所以$P(Y|X)$可以表示成$P(Y_i|X)$, $P(Y_i, Y_{i-1}|X)$的乘积形式。

根据指数概率模型和条件随机域的定义，对于链图G(V, E)，标注序列y在给定观测序列x下的概率可以正则化为如下一些式子的乘积[4]：

$$\exp(\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)). \tag{1}$$

其中$\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)$是$P(y_i,y_{i-1}|x)$对应的函数，$\sum_k \mu_k s_k(y_k,x,i)$是$P(y_i|x)$对应的函数，$t_j(.),s_k(.)$是特征函数，$\lambda_i,\mu_k$是通过训练得到的特征函数的参数。

特征函数定义为关于y,x的指示函数，每个特征函数给出一组与y,x的一个或多个特征相关的限制条件，当条件满足时，函数值为1，否则为0。

由于$s_k(y_i,x,i)$可以写成$s_k(y_i,y_{i-1},x,i)$，可以将t_j和s_k统一的具有$f_j(y_i,y_{i-1},x,i)$的形式。设

$F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x,i)$，那么

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp(\sum_j \lambda_j F_j(y,x)), \tag{2}$$

其中Z(x)是正则化因子，$p(y|x,\lambda)$为需要计算的条件概率。这里λ_j是待训练的参数，一般采用最大熵原则和GIS算法，具体参见文献 [1]。

2.3 与其他模型的比较

CRF是指数概率模型的一种，与其他模型相比，如传统概率模型[13], MEMM[14]等，其最大区别是CRF的指数模型刻画了整个序列的条件概率，由此解决了标注偏差问题。文[1]中针对标注偏差问题比较了CRF与MEMM，实验结果充分显示了CRF的优势。

CRF, HMM等模型也被称为序列分类模型。除了序列分类模型，更一般的分类模型和分类算法也可用于序列分类，比如SVM，Bayes分类器，神经网络等等，而这类模型和算法针对的是无序的向量数据，向量的分量之间不包含序列信息，没有直接的依赖关系。下表总结了序列分类器和一般分类器的一些区别：

表1 序列分类器和一般分类器的区别

Tab.1 Differences between sequence classifier and general classifier

	序列分类器	一般分类器
数据类型	有序或无序图	无序向量
适合问题	特定结构数据	向量数据
分类器实例	CRF，HMM，MEMM等	NB，SVM，NN等
常见应用领域	序列标注问题，词性标注，实体识别，DNA序列识别等	绝大部分分类问题，如文本分类，图像识别等

3 CRF 在时间信息识别的应用

要进行时间信息识别，首先应该明确什么是时间信息。简单地说，时间信息就是文档中用于表达一个时间点或一个时间段的词串，比如“1999-3-15”，“Friday”，“This afternoon”，“the day before yesterday”，“one year”等等。时间信息的定义标准有很多种，其中TIMEX [7]是被广泛采用的一种标准，在TIDES和ACE等比赛中被采用作为标准。在TIMEX识别问题中应用CRF，首先得将TIMEX识别问题转化为序列标注问题，我们考虑识别TIMEX串的起始位置和长度。对文档先进行切分和停用词过滤，得到的词按文档次序即构成了词序列，用标注B、I、O分别表示该词语位于某个TIMEX的起始、中间、外部，给每个词标上其对应的标注之后，所有的TIMEX的起始位置和长度就很明确了。比如句子“the seventeenth , in the afternoon..”的标注如下，而其中对应的TIMEX是“the seventeenth”和“the afternoon”。

句子：the seventeenth , in the afternoon .

标注：**B I O O B I** O

转化为标注问题之后，我们采用CRF模型进行标注，特征选取是标注过程中最关键的一步，特征的好坏对结果影响非常大，针对特定应用领域，应该选取最能反应领域知识的特征。一般常用的特征包括词形，小写形式，词性，词频等。由于TIMEX一般比较规范，一些明显的特征对于时间信息识别就非常有效，例如仅仅采用词形(小写形式)和词性作为特征就能正确地标注大部分词。添加其他有效的特征可以进一步提高结果，如词的组成(数字，字母等)，组成状态(大小写)，一些特殊的时间词(如月名，日名)等。下表是时间标注语料的一个样例，其中句子为“Very well, We'll see each other the ninth of August at two, or one. ”。其中的TIMEX为“ninth of August”，“two”，“one”。.表2中依次列出了每个词的词性，正确标注，CRF标注结果。

表2 标注语料样例

Tab.2 Labeling corpus sample

词	词性	正确标注	系统标注	词	词性	正确标注	系统标注
Very	RB	O	O	ninth	JJ	B	I
well	RB	O	O	Of	IN	I	I
,	,	O	O	August	NP	I	I
We	PP	O	O	At	IN	O	O

'll	MD	O	O	two	CD	B	B
see	VV	O	O	,	,	O	O
each	DT	O	O	Or	CC	O	O
other	JJ	O	O	one	CD	B	B
the	DT	O	B	.	SENT	O	O

CRF通过计算似然来选择候选特征集中的最优特征，本文的主要目的是对CRF与其他方法进行比较，不再具体介绍特征的选择过程，详细可参考[6]。

4 实验结果

实验使用语料来自TIDES Temporal Corpus，我们将其切分成两部分，其中训练语料约30000个词，测试语料6592词。我们选用了SVM方法进行标注，和CRF方法进行比较。采用如下的衡量指标：按词统计所有标注的正确率以及三个标注类别的精确率和召回率。具体公式如下：

$$\text{正确率} = \frac{\text{正确标注词数}}{\text{所有词数}}$$

$$\text{单类标注精确率(precision)} = \frac{\text{该类标注正确词数}}{\text{所有被标为该类的词数}} \quad (3)$$

$$\text{单类标注召回率(recall)} = \frac{\text{该类标注正确词数}}{\text{所有实际为该类标注的词数}}$$

4.1 实验结果

SVM是当前最热门的一种分类方法，我们用来和CRF进行对比实验。SVM采用的特征为词和词性，并使用了滑动窗口方法，窗口大小为3。对于CRF，在特征选择上我们采用两种特征选择方案，只用词（word）作为特征，以及用词(word)和词性(POS)作为特征。其中，表3和表4列出了SVM和CRF两种方案的结果，表3为它们的正确率，表4为单个类别标注的精确率和召回率。实验结果中CRF增加词性作为特征后比单用词作为特征多正确标注了18个词，实际上，使用词性后纠正了78个标注错误，但又产生了60个原本没有的错误。这说明每个特征都对标注分类有正面和反面的效果，如何选择特征是实际应用中的最重要问题之一。由于SVM没有考虑类似CRF的序列信息，使用相同特征的SVM，效果上要逊于CRF。

表3 三种方法的正确率

Tab.3 Total precision of three methods

	SVM	CRF(only word)	CRF(word+POS)
正确率	0.954	0.956	0.959

表4 单类标注的精确率和召回率

Tab.4 Precision and Recall of single label class

	Tag "B"		Tag "I"		Tag "O"	
	Precision	Recall	Precision	Recall	Precision	Recall
SVM	0.899	0.837	0.889	0.913	0.973	0.980
CRF(only word)	0.889	0.827	0.896	0.920	0.973	0.982
CRF(word+POS)	0.913	0.848	0.904	0.920	0.976	0.983

4.2 实验分析

CRF对于序列标注问题的处理，比其他分类方法具有较强的优势，然而由于语言的灵活性和统计方法本身的局限性，标注结果并非尽如人意。我们对标注错误按照不同词性进行了分类，以探讨不同词性在CRF的词标注的歧义度，分析其错误原因。下表是一些较重要类别的结果统计，使用的方法是CRF(word+POS)：

表5 不同词性的错误标注数统计

Tab.5 Label errors of different POS classes

	所有词数	SVM	CRF(word)	CRF(word+POS)
NN	751	67	55	49
PP	637	12	11	13
JJ	449	13	21	22
DT	565	55	59	52
NP	284	24	15	20
IN	717	54	52	46
CD	229	25	24	21

由上表可知，主要错误是由于NN，DT， IN 三种词性造成的，原因分别分析如下：

(1) NN：名词，最主要错误来自对数词如“eleven”“seventeenth”等的错分，因为这些词可能作为时间表达式的一部分，也可能表达其它数量或次序；还有比如“tomorrow”等时间表达式组成部分被标注为“O”，一部分因素是由于训练语料不充足而导致的漏分。

(2) IN：介词，大部分错误是将不是时间表达式组成的介词识别为时间表达式的前后缀或中间词，例如原文中的TIMEX“twelve”和“one”被识别为“twelve at one”；而少量错误表现为漏过了后缀或中间词，只有介词“that”出现了漏过前缀的情况。

(3) DT：the，a，an等冠词，主要问题在于这些词究竟是否作为时间表达式的起始，比如“ninth of August”和“the ninth of August”，到底“the”是否在TIMEX范围内还值得商榷，因此识别出来的短语并非完全错误。

特征选择是CRF识别中的一个重要问题，我们这里只作了简单的分析，通过这种分析，有利于进一步选择有效的特征，提高系统的性能。

5. 结论

条件随机域模型结合了条件概率模型的优势，并克服了标注偏差的问题。通过引入时间信息识别的研究，我们希望从理论和实验两方面说明CRF方法更适合于解决序列问题。本文特别比较了CRF和SVM等方法的差异，通过时间信息识别的实验证明，CRF更适合用于序列标注领域，具有很大的潜力。

CRF以其优良的性能得到越来越广泛的应用，与此同时，CRF本身也在不停地发展，产生了许多改进的模型，包括引入了核函数的CRF[10]，基于Bayes统计的CRF[11]，SVM和CRF模型相结合的最大间距马尔科夫模型(Max-Margin Markov Networks)[12]等。我们将来工作的重点将放在特征选取和CRF模型改进上。

参考文献：

[1] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning, 2001.

[2] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In Proceedings of Human Language Technology-NAACL, 2003.

[3] A. McCallum and W. Li. Early results for Named Entity Recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the 7th CoNLL, 2003.

[4] Hanna M. Wallach. Conditional Random Fields: An Introduction. Technical Report MS-CIS-04-21 University of Pennsylvania,2004.

[5] Thomas G. Dietterich. Machine Learning for Sequential Data: A Review. In Structural, Syntactic, and Statistical Pattern Recognition; Lecture Notes in Computer Science, vol.2396, T.Caelli(Ed.), pp. 15-30, Springer-Verlag, 2002.

[6] Andrew McCallum. Efficiently Inducing Features of Conditional Random Fields. In Proceedings of the 19 th Conference in Uncertainty in Artificial Intelligence (UAI-2003), 2003.

[7] Ferro L. , Gerber L. , Mani I. , Sundheim B. , And Wilson G. TIDES 2003 standard for the annotation of temporal expressions (2004). timex2.mitre.org

[8] David Ahn, Sisay Fissaha Adafre, Maarten de Rijke. Extracting Temporal Information from Open Domain Text: A Comparative Exploration. Journal of Digital Information Management, 3(1):14-20, 2005.

[9] Kadri Hacioglu, Ying Chen, Benjamin Douglas. Automatic Time Expression Labeling for English and Chinese Text, to appear in Proceedings of CICLing-2005, Mexico City-Mexico, Feb. 13-19, 2005.

[10] John Lafferty, Xiaojin Zhu, Yan Liu. Kernel Conditional Random Fields: Representation and Clique Selection. In Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004), 2004.

[11] Yuan Qi, Martin Szummer, Thomas P. Minka. Bayesian Conditional Random Fields. To appear in Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS 2005), 2005.

[12] Simon Lacoste-Julien. Combining SVM with graphical models for supervised classi¯cation: an introduction to Max-Margin Markov Networks.CS281A Project Report, UC Berkeley, 2003.

[13] Paz, A. (1971). *Introduction to probabilistic automata.* Academic Press.

[14] McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. *Proc. ICML 2000* (pp. 591–598). Stanford, California