软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(10):2340−2353 [doi: 10.3724/SP.J.1001.2013.04393]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

结构化学习的噪声可学习性分析及其应用*

于 墨, 赵铁军, 胡鹏龙, 郑德权

(哈尔滨工业大学 计算机科学与技术学院 语言语音教育部-微软重点实验室,黑龙江 哈尔滨 150001)

通讯作者: 于墨, E-mail: yumo@mtlab.hit.edu.cn

摘 要: 噪声可学习性理论指出,有监督学习方法的性能会受到训练样本标记噪声的严重影响.然而,已有相关理论研究仅针对二类分类问题.致力于探究结构化学习问题受噪声影响的规律性.首先,注意到在结构化学习问题中,标注数据的噪声会在训练过程中被放大,使得训练过程中标记样本的噪声率高于标记样本的错误率.传统的噪声可学习性理论并未考虑结构化学习中的这一现象,从而低估了问题的复杂性.从结构化学习问题的噪声放大现象出发,提出了新的结构化学习问题的噪声可学习性理论.在此基础上,提出了有效训练数据规模的概念,这一指标可用于在实践中描述噪声学习问题的数据质量,并进一步分析了实际应用中的结构化学习模型在高噪声环境下向低阶模型回退的情况.实验结果证明了该理论的正确性及其在跨语言映射和协同训练方法中的应用价值和指导意义.

关键词: 结构化学习;噪声 PAC 可学习性;词性标注;自然语言处理;协同训练;跨语言映射;半监督学习

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 于墨,赵铁军,胡鹏龙,郑德权.结构化学习的噪声可学习性分析及其应用.软件学报,2013,24(10):2340−2353. http://www.jos.org.cn/1000-9825/4393.htm

英文引用格式: Yu M, Zhao TJ, Hu PL, Zheng DQ. Theoretical analysis on structured learning with noisy data and its applications. Ruan Jian Xue Bao/Journal of Software, 2013,24(10):2340−2353 (in Chinese). http://www.jos.org.cn/1000-9825/4393.htm

Theoretical Analysis on Structured Learning with Noisy Data and its Applications

YU Mo, ZHAO Tie-Jun, HU Peng-Long, ZHENG De-Quan

(MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Corresponding author: YU Mo, E-mail: yumo@mtlab.hit.edu.cn

Abstract: Performance of supervised machine learning can be badly affected by noises of labeled data, as indicated by existing well studied theories on learning with noisy data. However these theories only focus on two-class classification problems. This paper studies the relation between noise examples and their effects on structured learning. Firstly, the paper founds that noise of labeled data increases in structured learning problems, leading to a higher noise rate in training procedure than on labeled data. Existing theories do not consider noise increament in structured learning, thus underestimate the complexities of learning problems. This paper provides a new theory on learning from noise data with structured predictions. Based on the theory, the concept of "effective size of training data" is proposed to describe the qualities of noisy training data sets in practice. The paper also analyzes the situations when structured learning models will go back to lower order ones in applications. Experimental results are given to confirm the correctness of these theories as well as their practical values on cross-lingual projection and co-training.

Key words: structured learning; PAC learning with noise; pos-tagging; natural language processing; co-training; cross-lingual projection; semi-supervised learning

有监督机器学习方法是机器学习领域最重要和最成熟的研究方向,并在如自然语言处理在内的很多具体

* 基金项目: 国家自然科学基金(61173073); 国家高技术研究发展计划(863)(2011AA01A207)
收稿时间: 2012-06-11; 修改时间: 2012-08-20; 定稿时间: 2013-02-04

任务中得到了广泛的应用.有监督学习方法的性能不仅依赖于标记样本的数量,也依赖于标记样本的质量.当标记样本中发生了错误,即当标记存在着噪声时,学习得到的分类器的精确率会受到影响.为了刻画数据噪声对有监督学习过程的影响,学习问题的噪声可学习性被提了出来[1,2].由于在实际应用中,无论样本的标记是自动获得的还是由人标记的,标记出现错误是难免的,因此,噪声可学习性理论作为无噪声环境下的经典学习理论的推广,对于实际应用有着重要的指导意义.另一方面,为了解决标注数据不足的问题,一些半监督学习方法被提了出来,其中,以协同训练为代表的一系列得到广泛应用的方法,其主要思想都是对无标记数据进行自动标记并将标记结果用于训练.而这些算法的有效性,则被建立在噪声可学习性理论的基础上.综上,在有监督学习问题中,噪声学习理论对于提高标记数据的质和量都有着重要的指导意义.

现有的噪声可学习性理论在应用上受到限制的原因之一是:这些理论都针对二类分类问题进行分析,然而实际应用中的问题往往并非二类分类问题,而是具有多个分类类别,甚至分类的目标是得到一种结构化的信息.具有结构化输出的在现实中具有广泛的存在性,比如自然语言处理中的词性标注/句法分析、图像处理领域中的物体识别以及生物信息学领域的基因片段识别和蛋白质结构预测,我们将这一类问题称为结构化学习问题.在对这些问题进行噪声可学习性的分析时,二类分类问题的噪声可学习性理论往往并不适用.

与理论发展不充分形成对比的是,结构化学习问题往往更容易遭遇到数据噪声的影响,因此对于噪声可学习性理论有着更大的需求.首先,结构化学习问题由于其标注的复杂性,对标注人员的专业知识有更高的需求,发生标注错误或标注过程中存在不一致性的情况也更容易发生,使得结构化学习在噪声环境下进行这一现象更为普遍;其次,由于结构化学习问题标注的困难性,可用于结构化学习的有标记数据往往较为稀少,从而导致结构化学习问题对半监督学习方法提出了更多的需求,因此需要相应的结构化学习的噪声学习理论对半监督学习方法进行指导.

本文主要关注这一类结构化学习问题在噪声环境下的学习行为.到目前为止,虽然研究人员们在结构化学习问题上开展了很多理论分析,但据我们所知,关于结构化学习的噪声可学习性理论的研究还是空白.在本文中,我们基于随机分类噪声假设[1],从结构化学习问题的噪声放大现象出发,对结构化学习问题的噪声可学习性进行了分析.噪声放大问题是指,在结构化学习中,标注数据的噪声会在训练过程中得到放大,使得训练过程中标记样本的噪声率高于标记样本的错误率.已有的噪声学习理论由于忽略了噪声放大现象,会倾向于低估噪声学习问题的复杂性.本文的工作首先把二类分类问题的噪声学习理论推广到多类上,通过将结构化学习问题简化为因子图[3]学习问题,并将因子图学习问题归纳为多类分类问题,从而实现了结构化学习问题的噪声可学习性的形式化分析.在此基础上,本文提出了有效训练数据规模的概念,用于描述噪声学习问题的数据质量.更进一步地,我们考虑实际应用中的结构化学习模型,对其在高噪声环境下向低阶模型回退的情况进行了研究.本文以词性标注为例,对上述理论进行了验证,说明本文的理论可以很好地对噪声环境下结构化学习方法的行为进行描述.

本文提出的结构化学习的噪声可学习性理论具有重要的应用意义,本文使用该理论对两个应用场景进行了分析.其中,

第 1 个应用是自然语言的跨语言映射问题中[4].

跨语言映射是把语言学标记从一种语言(源语言)映射到另一种语言(目标语言),从而为目标语言构造语料库资源的过程.以词性标注的跨语言映射为例,源语言中词汇的词性标记通过词对齐信息对目标语言的词进行词性标记.在映射得到的数据上,我们可以训练得到目标语言端的词性标注模型.由于双语的语法异构和词对齐错误的存在,跨语言映射所得到的标注数据存在着噪声.而本文的理论可以解释跨语言映射中使用一对一词对齐映射能够得到更好结果[4]的原因.

本文所选择的另一个应用是结构化学习中的协同训练方法[5].

协同训练方法是噪声学习理论在半监督学习中的直接应用,然而在结构化学习任务中,直接使用协同训练选择整个结构并非一种合理的选择,因为即使是较为准确的结构化标注数据之中也会存在一些标注不准确的子结构.为此,本文在基本的协同训练算法上改进得到了一种基于子结构选择的协同训练算法.这一改进的核心

思想基于前面的噪声可学习性理论,期望减少每轮迭代过程中新的训练数据噪声,并使其具有更高的有效数据规模.我们将这一方法应用于词性标注的跨语言映射任务中,其实验结果很好地符合了这一期望.

本文的主要贡献包括以下几个方面:

1) 本文描述了结构化学习问题中的噪声放大问题,并对其进行理论分析,得到了结构化学习问题的噪声可学习性理论(第 2 节).更进一步地,考虑到现实应用中的结构化学习模型往往是多阶模型的插值组合,本文将这一理论进行了扩展,从而对高阶结构化学习模型在噪声环境下向低阶模型退化的现象进行分析(第 3 节);
2) 本文提出两种实验方案对上述理论进行验证,实验结果表明,本文提出的理论能够更好地解释噪声环境下结构化学习的实验现象(第 4.4 节).本文的实验同时将跨语言映射问题与噪声可学习性问题建立了关联,这是据我们所知的首次从机器学习角度对跨语言映射问题进行的理论分析;
3) 基于本文所提出的理论,我们将协同训练方法在结构化学习问题上进行了推广和改进,这一方法的原理是:在每轮迭代中选取置信度最高的子序列,从而减小噪声放大问题对协同训练过程的影响.本文在词性标注的跨语言映射任务上对新算法的有效性进行了验证(第 4.5 节),说明该理论对于实际应用中的算法设计具有较好的指导意义.

1 相关工作

自从文献[1]提出噪声条件下二类分类问题的 PAC 可学习性理论以后,分类问题的噪声可学习性得到了研究人员的广泛关注,文献[6]证明了任何基于统计队列可学习的问题是噪声条件下可学习的.在此基础上的大部分工作[7,8]致力于探究如何获得随机分类噪声环境下的高效学习算法.近年来,研究人员们更多地关注训练样本具有的噪声并非随机分类噪声的情况,比如,文献[9−11]等工作对自然语言处理问题中的标注噪声及其对学习系统引入的固有偏差进行了研究.

大部分机器学习理论的工作都针对二类分类展开,由于现实中的问题往往不能简单地归纳为二类分类问题,一些研究人员对结构化学习问题的可学习性进行了研究,并证明了其推广能力的界.文献[12]在基于 0-1 损失函数的大边缘学习方法中,证明了线性模型的期望风险的界.文献[13]则证明,当损失函数为汉明距离时,大边缘学习方法的期望风险的界.文献[14]对上述两项工作中得到的推广能力的界进行了改进,并且对结构化学习优化算法的渐进一致性进行了分析.同时,与本文类似,文献[14]也对结构化输出分解成因子图的情况进行了分析.

尽管噪声可学习性和结构化学习问题的可学习性理论都得到了较为充分的发展,但据我们所知,当前在结构化学习问题的噪声可学习性理论方面还缺乏相关的研究.正如本文所说明的,由于在实际应用场景中复杂的结构标注更容易出现错误,在结构化学习问题上发展噪声可学习性研究是十分必要的.同时,结构化学习问题的噪声可学习性对半监督结构化学习算法的研究有着重要的指导作用.

2 结构化学习的噪声可学习性分析

2.1 结构化学习中的噪声放大问题

当训练数据带有噪声时,分类器的学习会变得比无噪声环境下的学习更加困难.一些研究人员已经证明,训练数据的噪声会对模型的推广能力产生影响,其中一个有代表性的工作为文献[1]在标准分类噪声上得到的结论.在二类分类问题中,噪声对于学习问题的推广能力的影响可以表示为如下定理:

定理 1(噪声数据的 PAC 可学习性)(Laird,1988). 若训练数据在采样过程中其标记具有为 η的噪声,则 L_*是假设空间中的最优假设(真实函数),L_i是在训练样本上得到的最优假设,且假设 L_i与 L_*在样本分布上输出不一致的概率为 $d(L_i,L_*)$.那么,当 $\Pr(d(L_i,L_*)\geqslant\varepsilon)<\delta$时,训练样本集合的大小 m 需满足:

$$m \geqslant \frac{2}{\varepsilon^2(1-2\eta_b)^2}\ln\left(\frac{2N}{\delta}\right) \tag{1}$$

其中,N 是假设空间的大小,η_b 是 η 的一个上界(关于 η_b 的确定,见文献[1]).

由定理 1 可以看到,当噪声 η<1/2 时,问题是噪声可学习的,且 η 越大,为了得到同样精度的分类器,理论上需要的样本就越多.

然而,当问题具有结构化输出时,数据标记的错误率与训练时的样本噪声两者并不完全等同.如图 1 所示,我们以标注具有噪声的链式条件随机场模型[15]为例来说明这一现象.

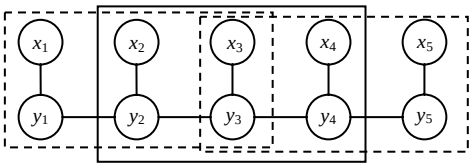


Fig.1 Enlargement of noises in chain CRF

图 1 链式条件随机场的噪声放大现象

图 1 中:$X=\{x_i\}$代表序列的输入;$Y=\{y_i\}$代表序列的标记,标记可能带有噪音.假设图中 y_3 对应的节点发生标记错误时,其他节点标注正确,则对于例子中这个长度为 5 的链,标记的错误率为 1/5.然而,当这一数据被用于训练一阶的条件随机场模型时,训练过程中的样本错误率可能不再是 1/5,而是有了一个“放大”.为了便于说明问题,我们将链式条件随机场的训练问题简化为以每个节点为单独样本的多类分类问题,对于每个节点 i,对应的分类任务是以$\{X,y_{i-1},y_{i+1}\}$为特征,对节点的标记 y_i 进行分类.从而在训练过程中,基于图 1 所示的结构可以为节点 i 构造训练数据$(y_i,\{X,y_{i-1},y_{i+1}\})$.从图 1 中可以构造 5 个这样的训练样本.

那么,在这 5 个训练样本上的标注错误率是多少呢?由于 y_3 是错误的标记,$(y_3,\{X,y_2,y_4\})$显然是标注错误的噪声样本.另一方面,考虑样本$(y_2,\{X,y_1,y_3\})$和$(y_4,\{X,y_3,y_5\})$,虽然 y_2 和 y_4 的标注是正确的,然而 y_3 存在的标注错误会导致样本的特征发生改变.不妨设第 3 个节点的真实标记为 y_3',则对节点 2,训练样本从 $(y_2,\{X,y_1,y_3'\})$ 变成了另一个样本$(y_2,\{X,y_1,y_3\})$,两个样本具有不同的特征,因此,原来正确的标记 y_2 对于这个新样本来说可能不再是正确的.更进一步地,样本$(y_2,\{X,y_1,y_3\})$在原始问题的数据分布中可能具有很小的概率或者概率为 0,从而为问题的解决提供了无用的甚至是错误的样本.同理,节点 4 也会出现这一情况.由此,在 1 阶模型中,若一个链式结构中有一个节点被标记错误,则当这一链式结构被用作训练数据时,受到这个错误标记影响的节点最多可能有 3 个.在最坏情况下,这 3 个节点都构成错误样本.从而在上述简化的多类分类问题中,样本的噪声被放大为 3/5.我们将这一现象称为结构化学习的噪声放大问题.

综上,在结构化学习中,当标注数据带有噪声时,这一噪声会在训练过程中被放大,从而使得训练过程中标记样本的噪声率高于标记样本的错误率.在链式条件随机场的例子中,这一噪声会被放大到 η~3η 之间.结构化学习的噪声放大问题说明,定理 1 形式的噪声可学习性理论低估了数据标记的噪声对训练过程产生的影响.为了形式化地刻画噪声放大问题,从而为结构化学习的噪声可学习性建立更合适的理论,在下一节中,我们将考虑链式的生成式分类模型,在这一相对简化的问题上对结构化输出问题的噪声可学习性进行分析,并给出基于 PAC 的模型推广能力的界.注意,在本文中,我们只使用 PAC 理论为例进行分析,对于无穷假设集合,可以使用文献[6]的方式将类似的证明过程推广到 VC 理论[16]上.

2.2 基于因子图的结构化输出问题的噪声可学习性分析

二类分类问题的噪声可学习性理论可以较容易地推广到多类情况(见下文定理 2 的证明),然而将这一理论推广到结构化学习问题并不直接.最大的困难在于:结构化学习问题的类别集合不固定,直接把模型的整个输出看作数据的类别标记会遭遇一些困难.比如在序列标注问题中,一个长度为 L 的序列可能的标记方式有 C^L 种.因为这一问题的类别数与序列长度相关,而输入序列的长度不固定,从而使得结构化学习问题难以带入到定理

1 的框架中进行分析.

为了解决这一问题,我们考虑把输出结构拆分成子结构进行分析,而不是简单地看作一个整体.在这里,我们考虑基于概率图模型的结构化学习问题,将模型表示为因子图[3],并以因子为最小单位进行分析.图 2 所示为图 1 中链式条件随机场对应的因子图,其中,“□”代表因子,与因子向量相连的节点为因子的变量.

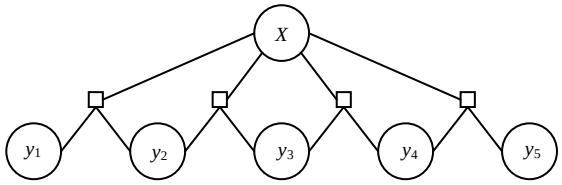


Fig.2 Factor graph of sequential labeling models

图 2 序列标注模型的因子图表示

在因子图中,整个图模型的概率可以根据因子图进行分解,并表示为正比于各因子势函数的乘积.在链式条件随机场中,模型的参数具有时序(位置)不变性,即不同位置的因子共用同一套参数表示.因此对于上述因子图,其联合概率可以表示为

$$P(X,Y)=\exp\{\sum_{i=1}^{L(X)}\psi(y_i,y_{i+1},X)\}\Big/Z=\exp\{\sum_{i=1}^{L(X)}\sum_j\lambda_j f_j(y_i,y_{i+1},X)\}\Big/Z \tag{2}$$

其中,$\exp\{\psi(y_i,y_{i+1},X)\}$是节点$(y_i,y_{i+1},X)$对应因子的势函数,$f_j$是定义在因子上的特征,$\lambda_j$是特征$f_j$的取值,$L(X)$是序列长度.可见,在序列标注的概率模型中,模型的参数只描述链上的因子,因此,模型空间中的每个假设可以唯一地表示为对应的因子图参数$\{\lambda_j\}$.从而基于因子图,我们为生成式结构化学习问题的模型参数和因子图的参数建立起了一一对应的关系,将结构化输出任务的可学习性转化为因子图参数的可学习性.

更进一步地,我们考虑生成式模型的优化目标和因子图的优化目标之间的关系:对于生成式模型,在训练时对训练集 D 进行极大似然估计,其优化目标为

$$\arg\max_\theta\sum_{(X,Y)\in D}\log(P(X,Y\mid\theta))=\sum_{(X,Y)\in D}(\sum_{i=1}^{L(X)}\log(\psi(y_i,y_{i+1},X;\theta))-\log Z_{L(X)}(\theta))=\sum_{\psi}\log(\psi(y,y',X;\theta))-R(\theta).$$

这一目标可以看作是对训练数据中的所有因子图进行极大似然,由于 $R(\theta)=\sum_{(X,Y)\in D}-\log Z_{L(X)}(\theta)$ 与训练数据的取值无关,因此可以看作一种基于结构化信息的正则项.当$Z(\theta)$=1(如 HMM)或$Z(\theta)$可以分解到因子图(如局部归一化模型),即与θ无关时,这一优化目标等价于以每个因子ψ为独立样本进行极大似然的结果.此时,对于生成式模型,因子图的结构化学习问题的优化目标与以因子为基本元素的学习问题等价,本文的理论主要考虑这种情况.当 $Z(\theta)$与θ相关且不可分解时,本文的理论只能得到近似的结果.这是因为本文基于 PAC 可学习性,相当于经验风险最小化,不包含正则项对应的结构风险.我们将在后续工作中把理论推广到 VC 理论中,并对这一现象进行精确分析.

基于上述分析,结构化输出问题的核心是如何估计因子的势函数.一旦因子图的参数估计正确,那么结构化学习的模型也就得到了正确的估计.因此,我们可以将结构化学习模型的可学习性简化为模型因子图上势函数的参数可学习性.为了描述因子图的可学习性,我们首先引入如下定理:

定理 2(多类分类任务的噪声 PAC 可学习性). 若训练数据在采样过程中其标记具有为η的噪声,且假设有噪声样本标记成任何一个其他类别的概率相同,L_*是假设空间中的最优假设,L_i 是在训练样本上得到的最优假设,且假设 L_i 与 L_*在样本分布上输出不一致的概率为 $d(L_i,L_*)$.那么,当 $\Pr(d(L_i,L_*)\geqslant\varepsilon)<\delta$时,训练样本集合的大小 m 需满足:

$$m\geqslant\frac{2}{\varepsilon^2\left(1-\dfrac{C}{C-1}\eta_b\right)^2}\ln\left(\frac{2N}{\delta}\right) \tag{3}$$

其中,N 是假设空间的大小,η_b 是η的一个上界.

其证明可见附录 1.

这一定理的结论与多类分类问题的弱分类器的直观定义相符合.即,在类别平衡的数据上,多类分类问题的弱分类器应满足分类的准确率大于 1/C,对应的错误率$\eta<(C-1)/C$.基于定理 2 的结论,如果我们将因子图的势函数学习问题看成是一个因子图上的多类分类问题,所有输出节点可能的输出组合构成了类别集合,则可以得到如下定理 3.

定理 3(生成式 1 阶序列标注模型势函数的噪声 PAC 可学习性). 对于图 2 所示的因子图,假设每个输出节点具有 C 个类别.若训练数据在采样过程中其标记具有为η的噪声,且假设有噪声样本标记成任何一个其他类别的概率相同,$L_*=\{\lambda_{*j}\}$是假设空间中的最优假设,$L_i=\{\lambda_{ij}\}$是在训练样本上得到的最优假设,且假设 L_i 与 L_*在样本分布上输出不一致的概率为 $d(L_i,L_*)$.那么,当 $\Pr(d(L_i,L_*)\geqslant\varepsilon)<\delta$时,训练样本集合包含因子数 m 需满足:

$$m\geqslant\frac{2}{\varepsilon^2\left(1-\frac{C^2}{C^2-1}(1-(1-\eta_b)^2)\right)^2}\ln\left(\frac{2N}{\delta}\right) \tag{4}$$

其中,N 是假设空间的大小,η_b 是η的一个上界.

证明:每个因子包含两个输出节点,我们将每个因子看成是一个样本,从而可以将因子图的分类任务看成是一个多类分类问题 $p(y_1,y_2|x)$,其类别对应着两个输出节点的输出(y_1,y_2),因此,可能的类别数为 C^2 个.

另一方面,每个节点的错误率为η,则整个势函数的类别标注错误的概率为两个节点任意一个标错或者全标错的概率为 $1-(1-\eta)^2$.

将类别数 C^2 和噪声 $1-(1-\eta)^2$ 代入公式(2),从而定理 3 得证. □

由于当$\eta<(C-1)/C$ 时,有 $\frac{C^2}{C^2-1}(1-(1-\eta_b)^2)>\frac{C}{C-1}\eta_b$;另一方面,当$\eta\geqslant(C-1)/C$ 时,多类分类问题和结构化问题都是不可学习的.因此,定理 3 说明,在标记错误率相同的有噪声数据上,只要问题是 PAC 可学习的,则为了使模型获得同样的准确率,具有链式结构输出的问题与每个节点具有同样类别数的多类分类问题相比,对样本有更多的需求.

基于定理 3 的证明,我们可以得到下面的定理 4:

定理 4(生成式 d 阶序列标注模型势函数的噪声 PAC 可学习性). 在定理 3 的条件下,当使用 d 阶序列标注模型时,训练样本集合包含因子数 m 需满足:

$$m\geqslant\frac{2}{\varepsilon^2\left(1-\frac{C^{d+1}}{C^{d+1}-1}(1-(1-\eta_b)^{d+1})\right)^2}\ln\left(\frac{2N}{\delta}\right) \tag{5}$$

其中,N 是假设空间的大小,η_b 是η的一个上界.注意到,随着子结构变得复杂,整个子结构的标注噪声$(1-(1-\eta_b)^{d+1})$会随之上升,从而导致公式(5)不等号右边随着 d 的增大而增大,即模型学习对样本的需求量有所增加.同时,因子的结构越复杂,从一条链中可以抽取的有标记因子也越少(为链的长度 $L-d$ 个),使得当样本有限时,高阶模型的学习更加困难,甚至不能满足可学习性要求.

定理 3 和定理 4 中对噪声的分析只是理论上的简化情况,因为在实际应用中,同一因子相邻类别节点是否发生错误往往不会是条件独立的.然而,当把因子的所有输出节点当作一个整体时,如果能够得到这个整体的标记错误率η_d,则可以用公式(6)替换定理 4 中的公式(5):

$$m\geqslant\frac{2}{\varepsilon^2\left(1-\frac{C^{d+1}}{C^{d+1}-1}(1-\eta_d)\right)^2}\ln\left(\frac{2N}{\delta}\right) \tag{6}$$

2.3 有效训练数据规模

为了更直接地对训练噪声和分类器精度的关系进行表示,我们引入有效训练数据规模这一概念.观察公式(5),不等式右边与除去分母中与噪声相关的多项式,形式与 PAC 可学习性的结论相同.而分母的噪声项则导致

为了达到同样的推广能力,噪声学习相比无噪声学习,需要更多的训练数据.这也意味着,在噪声学习的设置下,训练数据对分类模型起到的效用由于噪声的存在,要比无噪声学习设置条件下更小.因此,有噪声训练数据与同样规模的无噪声数据相比对学习过程来说“价值更低”.我们定义有效训练数据规模 m' 如下:

定义 5(噪声学习的有效训练数据规模). 在定理 4 与公式(6)的条件下,定义噪声学习的有效训练数据规模如下:

$$m'=\left(1-\frac{C^{d+1}}{C^{d+1}-1}(1-\eta_d)\right)^2 \tag{7}$$

基于上述定义,公式(5)可表示为

$$m'\geqslant\frac{2}{\varepsilon^2}\ln\left(\frac{2N}{\delta}\right) \tag{8}$$

通过引入有效训练数据规模这一定义,可以清楚地看到,当训练数据有噪声时,对于同一模型,学习的推广能力仅依赖于有效数据规模的大小,因此,定义 5 可以表示为:在具有相同噪声的数据上进行学习时,对于同一模型,有效训练数据规模越大,模型的性能就越好.

3 实际应用中的模型回退现象分析

定理 4 说明,在样本噪声较大的情况下,随着模型阶数的增加,结构化噪声会随之急剧增大,导致同样的训练集合所对应的有效训练样本规模急剧缩小.然而在实际应用中,这一现象往往不会出现,对应于定理 4,事实上是对实际噪声放大现象的高估.这是因为在实际应用中,因子的势函数 $\psi(y_i,y_{i+1},X;\theta)$ 往往可以分解为不同阶数的势函数的乘积.在 HMM 中,表现为转移概率和发射概率;在 1 阶 CRF 中,表现为两类不同的特征 $f_i(y_i,y_{i+1},X)$ 和 $g_i(y_i,X)$.从而,结构化问题的模型可以被看作高阶模型和低阶模型的组合.需注意,在实际应用中,为了使模型能够正常工作,即使模型不会明确对势函数进行拆分,这种回退往往也是不可避免的.比如,自然语言处理任务中常用的模型插值平滑算法就是在高阶模型处理能力受限时,模型向低阶的一种回退.

实际应用中,势函数的这一拆分使得当样本噪声较大时,高阶模型由于噪声具有更急剧的放大,相比低阶模型,更不可信,因此,模型中高阶模型的权重也随之降低,使得模型“回退”到低阶的情况.举例而言,对于二阶模型 HMM,当噪声较大导致标记间的三元特征不可信时(具有较为平滑的分布),若二元特征仍然具有较好的判别性,则模型的 1 阶部分仍可以较好地工作,因此模型的行为会偏向 1 阶模型.

在绝大多数应用任务中,标记数据的噪声都不会过高,从而单独使用定理 3 和定理 4 可以合理地描述模型在噪声数据上学习的行为.然而,研究高阶模型在噪声数据上的回退行为,对于实际应用中高阶模型的建立、模型阶数的选择及其行为分析(比如自然语言中常见的语言模型插值以及句法分析的平滑)都具有指导作用.

我们可以将噪声条件下高阶模型的回退行为形式化地加以表示:

为了简化表述,我们将 d 阶的因子图的势函数表示为 $\psi_d(y,y_{-1},\ldots,y_{-d},X)$(简记为 ψ_d),则 $\ln\psi_d$ 可以看作是多个不同阶数势函数 $\phi_i(y,y_{-1},\ldots,y_{-i},X)$ 的线性插值结果,即 $\ln\psi_d=\sum_{i=0}^{d}\ln\phi_i$.

在同一训练集 D 上,对于 d 阶模型,可以得到训练过程的经验风险 $E_d(D)$,以及通过经验风险最小化得到的模型势函数表示 $\psi_d(D)$.若此时有 $E_{d-1}(D)\sim E_d(D)$,即 $d-1$ 阶模型得到的经验风险与 d 阶模型接近,则 $\psi_{d-1}(D)\sim\psi_d(D)$,即 $\phi_d(D)\to 0$.此时,d 阶模型的参数与 $d-1$ 阶类似,即 d 阶模型向 $d-1$ 阶发生了回退.

在训练数据带有噪声的结构化学习任务中,阶数越高,噪声放大现象越明显,相应的 d 阶特征判别能力越接近随机,从而有 $\phi_d(D)\to 0$,因此,模型更容易向低阶回退.当噪声很大导致标记间的上下文信息完全不可信时,模型最终会退化到 0 阶,结构化学习问题退化到多类分类问题.此时,可以使用本文提出的定理 2,即不考虑噪声放大的情况,对结构化学习问题的噪声可学习性进行分析.

在无噪声的学习任务中,高阶模型可能会过拟合训练数据,从而总会导致更小的经验风险,同时具有更大的置信风险.而低阶模型的表达能力较弱,具有更高的偏差,从而具有较大的经验风险.因此在模型选择时,需要对模型的两种风险进行一个权衡,即考虑模型的结构风险(structural risk)[16].结构风险是统计学习理论中基于有限

样本集对整个样本分布上的期望风险所作的估计.当将同样的思想推广到有噪声学习任务中时,需要注意到高阶模型不仅具有更复杂的假设空间,还会遭受更严重的噪声放大,从而其置信风险会进一步增大.

根据上述讨论以及定理 4,我们可以在噪声环境下对结构化学习任务的结构风险进行最小化(SRM).给定训练集规模 m,如果我们可以对数据标记的噪声进行估计,则可以得到 d 阶模型的结构风险为 $S_d(D)$:

$$S_d(D) \leqslant E_d(D) + 2\sqrt{\frac{2}{m\left(1-\frac{C^{d+1}}{C^{d+1}-1}(1-(1-\eta_b)^{d+1})\right)^2}\ln\left(\frac{2N_d}{\delta}\right)},$$

其中,N_d 是 d 阶模型假设空间大小.从而在噪声环境下对结构化学习模型进行模型选择时,可以根据结构风险最小化理论,选择模型的阶数 $d=\mathrm{argmax}_i S_i(D)$.

4 实 验

4.1 实验设置

本节以使用生成式的隐马尔可夫模型(HMM)来训练的词性标注为例,验证第 2 节中提出的结构化学习问题的噪声可学习性理论是否能够更好地符合现实中的实验现象.为了观察模型阶数对噪声放大现象的影响,我们分别训练了一阶 HMM 和二阶 HMM,并对其性能进行比较.一阶和二阶的 HMM 的有向概率图模型及其因子图如图 3 所示.实验中,我们使用 Hunpos[17]对模型进行训练和解码.

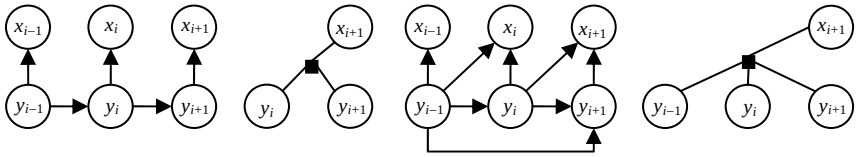


Fig.3 Directed probabilistic graphs and factor graphs of first-order (left) and second-order (right) HMMs

图 3 1 阶(左 1)和 2 阶(左 3)HMM 的有向概率图及其因子图(右侧)表示

我们在以下 3 个任务上对结构化学习的噪声可学习性理论进行了实验验证:

(1) 汉语词性标注的随机噪声实验

以宾州中文树库[18](CTB)第 1 节~第 815 节以及第 1001 节~第 1136 节为训练数据,第 816 节~第 885 节以及第 1137 节~第 1147 节为测试数据.训练数据包含 15 620 个句子、403 070 个词.我们在训练数据上添加为 η 的随机噪声,具体做法是:对于每个词,以 η 的概率改变其词性,且改变到其他词性的概率为均布.

(2) 基于跨语言映射的词性标注

实验中,我们以词对齐信息作为桥梁,使用源语言的词信息对目标语言的词进行词性标记,并在得到的标记数据上训练目标语言的词性标注器.选择这一实际应用作为验证的原因是:跨语言映射问题在一定程度上可以满足定理 3 的随机噪声假设——一个目标语言句子的跨语言映射结果标注错误与否取决于相应的源语言句子,而给定目标语言句子,对应的源语言翻译不是唯一确定的,而是具有随机性.对于同一个目标语言句子中的词,跨语言映射标注的正确与否也因此具有随机性.

(3) 结构化输出问题的协同训练算法

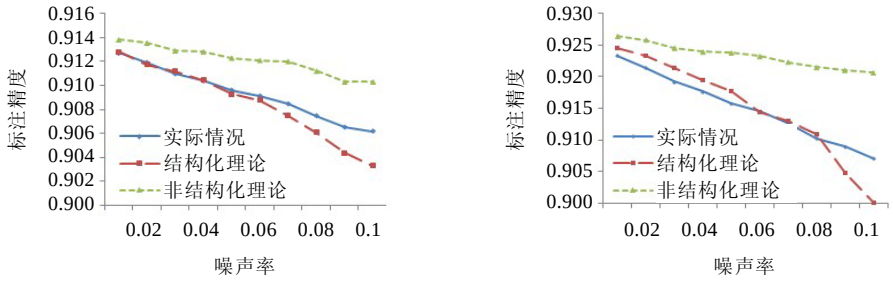
我们在词性标注的跨语言映射任务中,对结构化输出问题的协同训练算法进行改进,并分析其受到噪声的影响.协同训练是应用最为广泛的半监督学习方法之一,这一方法将问题的样本特征集合划分为两个视角.在每个视角上,协同训练方法学习一个分类器,用于对无标记数据进行标记并提供给另一个视角的分类器进行学习.协同训练的有效性基于机器学习模型的噪声可学习性:当一个分类器标注了一些无标记数据给另一个分类器用作训练时,另一个分类器相当于在有噪声的训练数据上进行学习.因此,结构化输出问题的协同训练可以作为本文理论的一个很自然的验证场景.

4.2 汉语词性标注的随机噪声实验

在本节中,我们在具有随机噪声的汉语词性标注任务上对本文提出的噪声放大现象和高阶模型回退现象进行验证.首先,考虑在绝大多数的应用任务中,标记数据的噪声一般不会过高,我们基于第 4.1 节所描述的有噪声的 CTB 数据,获取类别标记错误率为 0~0.1 范围内的噪声数据,观察模型在其上的学习行为,从而验证理论在一般应用场景下的正确性.对于同一噪声率,我们随机产生 100 组噪声数据,在每组数据上,训练词性标注器并在测试语料中进行测试,获取标注精度的平均值.

由于噪声学习理论无法直接对模型精度进行预测,为了衡量理论是否能够更好地解释实验现象,我们通过如下方法获得在某一噪声数据集上学习到的模型的理论性能:对于每一个噪声率,根据训练集的规模和噪声大小,可以根据公式(7)得到训练集的有效训练数据规模 m',并从无噪声的 CTB 训练集中随机获取规模为 m'的数据训练模型(在后文中称为有效无噪声训练集),并得到另一组分类精度.通过比较这组分类精度与在噪声数据上学习得到的分类器精度,可以观察理论与实验的拟合情况.

图 4 描述了在噪声较小的情况下,取 d=1 和 d=2 时定理 4 与实际噪声数据上实验结果的拟合程度.图中实线代表在噪声数据上训练得到的词性标注器的精度.我们根据公式(7)在相应噪声下进行计算得到有效无噪声训练集,在其上训练得到的词性标注器精度用长虚线表示.为了方便对比,我们根据定理 2 的计算结果得到相应的有效无噪声训练集,这一计算结果对应着传统的不考虑噪声放大问题的理论性结论,相当于公式(7)中 d=0 的退化情况,其上的词性标注器精度用短虚线表示.实验结果表明,在噪声率较小的情况下,在噪声数据上训练得到的分类器的实际精度与定理 4 的预测结果很接近.说明定理 4 计算得到的有效训练数据规模与实际情况相符,能够较为准确地刻画结构化学习中的噪声放大现象.另一方面,定理 2 倾向于高估分类器的精度,因为定理 2 不考虑噪声放大问题,从而会对噪声数据的质量进行高估(即高估噪声数据的有效训练规模).用另一种说法,即学习理论在不考虑噪声放大问题时,会倾向于低估噪声学习问题的难度.注意到,当噪声变大时,本文所提出理论的估计结果会偏低,这在一定程度上是因为词性标注任务中无法观测到所有的特征和标记组合(即词和词性的组合),当有效数据规模一致时,有噪声数据要比无噪声数据规模更大,从而包含更多的词.而新理论对应的有效无噪声训练集规模较小,因此在测试集上的未登录词的比例较大.理论估计结果比真实的有噪声训练结果偏低的另一个原因是模型的回退,见第 4.3 节.



(a) 1 阶模型 (b) 2 阶模型

Fig.4 Relation between the accuracies of tagging and noise rates in chain CRF

图 4 链式条件随机场标注精度和噪声率的关系图

4.3 高噪声条件下的模型回退实验

由第 3 节所述,在实际应用中,随着噪声的增加,高阶模型会变得不可信赖并向低阶模型回退.为了对这一现象进行研究,我们获取错误率在 0~0.2 范围内的噪声数据,并在其上对较高的噪声环境中模型的实际学习行为与理论的符合程度进行了分析.同样使用理论对应的有效无噪声训练集获得分类精度理论的预测结果.这里,我们以 2 阶模型为例在噪声数据上学习词性标注器.根据第 3 节的分析,在实际应用中,模型的性能应该与相应噪声条件下性能最好的低阶模型接近.我们通过本文中的理论,对同一噪声数据上不同阶数模型的理论性能进行

估计,并期望实际结果与最优的低阶模型相近.估计理论性能的具体做法是:对于每一阶数 d,根据理论计算出 d 阶情况下噪声数据对应的有效训练数据规模,并在相应的有效无噪声训练集上训练 d 阶模型.具体地,我们使用定义 5、d=2 的情况来估计 2 阶模型的性能(长虚线),使用定理 3 估计 1 阶模型的性能(点划线).因为实际情况下使用 2 阶模型,因此在噪声较小时,其性能会与 2 阶的理论结果保持一致;当噪声增大时发生回退,会趋近于 1 阶模型的结果,然而此时 2 阶模型所提供的信息并没有被完全抛弃.由于 2 阶模型使用更为复杂的特征,即使在高噪声下,这些特征与类别标记仍然具有一定的相关性(随着噪声的上升,相关性会降低),因此,模型在主要依赖于更具判别性的 1 阶特征的同时,仍然会利用一些 2 阶特征的有效信息,使得实际结果仍会略高于 1 阶模型的结果,图中曲线也符合这一推断.从图 5 可以看到,当噪声较小时(η<0.1),2 阶模型性能更优,对应着实际模型的表现与 2 阶模型的理论估计结果更为接近;当噪声较大时,实际模型的性能与定理所估计的 1 阶模型的性能相近,说明模型回退到 1 阶模型的情况.图中高噪声下 2 阶理论结果急剧下降,这在一定程度上是由于受到第 4.2 节中所述的未登录词的影响.实验结果表明,基于本文提出的定理 3 和定理 4,我们可以与实际情况很一致地刻画第 3 节中的高阶模型回退过程.这一结果同时也进一步证明了定理 3 和定理 4 关于有效训练数据规模描述的正确性.

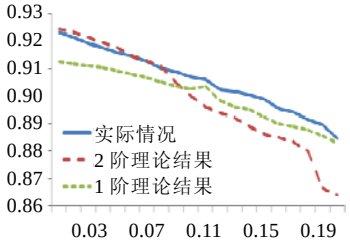


Fig.5 Back-Off from 2-order model to 1-order model for large noise rates

图 5 高噪声下 2 阶模型向 1 阶模型的回退

4.4 跨语言映射上的噪声放大问题分析

根据定理 3 和定理 4,通过使用有效数据规模对样本和噪声两者的综合作用进行描述,我们可以对分类器在噪声数据上训练后的准确率进行定性描述.在本节中,我们将通过对跨语言映射过程的错误率、映射结果用于训练时的训练噪声以及目标语言端分类器的词性标注精度三者之间的关系进行考察,对这一定性描述的准确性进行验证,从而进一步证明定理的正确性.

实验中使用 FBIS(LDC2003E14)作为双语语料,进行从英语到汉语的词性标注跨语言映射,并使用与上一任务相同的 CTB 测试集进行测试.我们使用文献[19]中的方法,以 CTB 第 1 节~第 20 节(237 句)作为汉语端少量有标记数据,在其上统计得到双语词性的映射规则.英语端使用斯坦福词性标注器[20]进行标注.由于对齐语料没有标准的词性标注,直接在其上观察映射错误率和噪声放大较为困难.为了解决这一问题,我们在 FBIS 语料中混入了 4 173 句 CTB 训练集语料及其对应的英文翻译结果(LDC2003E07,1-315),并重新得到一组词对齐的结果.由于新加入的对齐语料规模较小,我们认为其对对齐结果的影响较小,因而仍然能够反映 FBIS 语料的词对齐情况.经过跨语言映射过程,这些新加入语料的汉语端得到了从英语端映射的标注,通过观察其上的标注错误率和噪声放大情况,可以对 FBIS 上的相应信息进行估计.注意,在训练目标语言端分类器时,仍然只使用 FBIS 的对齐结果,这些新加入的数据不会产生影响.

实验中使用 GIZA++[21]获得双语语料的词对齐.由于在 GIZA++的输出结果中,一个目标语言的汉语词可能对应到英语端的一个词,也可能对应到多个词甚至对空,我们采用如下两种方式对 GIZA++的输出进行处理:

- 所有对齐:我们使用文献[19]中的方法,对词的对齐结果进行处理,将汉语词的一对多对齐处理成一对一对齐,并将对空的汉语词的词性标为名词 NN;
- 一对一对齐:简单地获取所有只对齐到英语端一个词的汉语词构成的子串.

上述两种方法处理后的对齐结果对应着两种不同规模噪声的训练集,其对应的映射结果错误率见表 1.可见,只使用 1-1 对齐得到的映射结果比使用所有对齐具有更好的精度(5.89%).

Table 1 Impacts of word alignments with different noise rates on projected data set

表 1 具有不同噪声率的词对齐结果对映射结果的影响

词对齐类型	覆盖词数	跨语言映射噪声率(%)
所有对齐	99 934	17.33
1-1 对齐	72 155	11.44

另一方面,根据 HMM 的因子图表示,可以在 CTB 双语语料汉语端的映射结果上分别得到 1 阶和 2 阶 HMM 的因子输出节点错误率(即相应 HMM 训练中的噪声放大结果),从而对整个对齐语料映射结果的噪声放大和有效数据规模进行估计.根据定理 4 和公式(7),模型的精度与有效数据规模呈正相关.基于上述设置,我们对这一结论进行了验证,结果见表 2.其中,FBIS 数据的训练数据噪声和有效训练数据规模均为使用 CTB 双语语料上的噪声放大结果得到的估计值.注意到,由于词的重复率和语料领域差异问题,这里的有效数据规模无法与前面 CTB 的结果进行直接比较.然而,不同的跨语言映射方法对应的有效训练数据规模相对大小的比较是有意义的.

Table 2 Relations between accuracies and noise rates in cross-lingual projection tasks (CTB tag set)

表 2 跨语言映射任务中词性标注精度与训练数据噪声的关系(CTB 词性标注集合)

模型阶数	词对齐类型	训练数据规模	因子图噪声(%)	有效训练数据规模	标注精度(%)
1 阶	所有对齐	7 053 052	31.33	3 322 700	78.79 (具体为 78.7876)
	1-1 对齐	5 398 197	21.20	3 350 163	78.79 (具体为 78.7896)
2 阶	所有对齐	7 053 052	42.79	2 308 179	78.67
	1-1 对齐	5 398 197	29.55	2 678 869	79.55

实验结果表明,尽管使用所有对齐得到的训练数据规模较大,然而其上本来较高的映射错误率在结构化学习的训练过程中被进一步放大,最终导致基于所有对齐得到的有效训练数据规模反而较小.相应地,在 1 阶和 2 阶模型中,基于一对一对齐的分类器在 CTB 的测试集上也获得了较高的分类精度.这与定理 4 的结论是相符的.更进一步地,对于 1 阶模型,基于所有对齐得到的有效训练规模与基于一对一对齐得到的有效数据规模的差异不大(3322700/3350163=99.18%),相比之下,2 阶模型上对应的比值(2308179/2678869=86.16%)要大得多.根据定理 4,这对应着 2 阶模型上,基于两种对齐训练得到的分类器的性能差也应相比 1 阶模型更为明显,而实验结论也证明了这一点.实验中,2 阶模型的性能差为 0.88%,而 1 阶模型的性能差仅为 0.002%,几乎可以忽略不计.

为了进一步为定理 4 的正确性提供佐证,在目标语言端,我们使用通用词性标注集合[22]代替 CTB 风格的词性标注定义,并将跨语言映射的结果和测试数据都通过文献[22]一文中提供的映射表转换为通用词性标注集合.由于通用词性标注集合相比于 CTB,其词性的定义分类粒度较大(只对较大粒度的名词、动词、形容词等进行),因此跨语言映射的准确率会有一定提升,从而使得在其上的实验具有较低的噪声.

我们在通用词性标注集合上重新进行了上述实验,结果见表 3.

Table 3 Relations between accuracies and noise rates in cross-lingual projection tasks (universal tag set)

表 3 跨语言映射任务中词性标注精度与训练数据噪声的关系(通用词性标注集合)

模型阶数	词对齐类型	训练数据规模	因子图噪声(%)	有效训练数据规模	标注精度(%)
1 阶	所有对齐	7 053 052	21.53	4 325 871	85.26
	1-1 对齐	5 398 197	14.79	3 909 804	84.78
2 阶	所有对齐	7 053 052	30.36	3 418 611	85.27
	1-1 对齐	5 398 197	20.85	3 380 859	85.10

可以看到,由于通用词性标注集合上训练数据噪声普遍较低,即使在结构化数据上噪声得到了放大,无论对 1 阶模型还是 2 阶模型,使用所有对齐数据得到的有效训练数据规模均高于一对一对齐的有效训练数据规模.相应地,对两种模型,使用所有对齐数据进行跨语言映射得到的分类器与基于一对一对齐的分类器相比,均具有较高的精度.注意到,在通用词性标注集合上,1 阶模型和 2 阶模型在所有对齐数据上得到的分类器性能相当.然

而对于 1 阶模型,一对一对齐得到的有效训练规模与通过所有对齐得到的有效数据规模的比值(3909804/4325871=90.38%)相比于 2 阶模型上对应的比值(3380859/3418611=98.90%)要小得多,因此,基于一对一对齐的 1 阶模型与基于所有对齐的 1 阶模型的性能差也相比于 2 阶模型更大(0.48% vs. 0.17%).

上述实验说明了在跨语言映射这一实际问题中,训练数据规模、训练数据噪声与分类精度三者之间的关联与定理 4 的一致性,从而证明定理 4 可以较为准确地描述结构化学习问题在有噪声情况下的行为.

4.5 结构化协同训练的跨语言映射实验

本文的结构化学习的噪声可学习性分析可以直接用于解释协同训练方法的实验现象以及对它的改进.本节以协同训练的数据选择为例证明本文理论的有效性.

按照第 4.4 节中的设置,我们将上述基于子序列的协同训练方法应用到词性标注的跨语言映射任务中.使用第 4.1 节所述的少量目标语言数据训练初始的目标语言模型,并使用第 4.4 节中具有一对一词对齐的对齐数据构建初始的映射模型.在实验中,采用与文献[19]同样的设置,以跨语言映射作为一个视角,以目标语言端单语模型作为另一个视角.每轮迭代有 50 000 个句子被更新,直到所有的句子被更新一遍.在分类器输出的置信度确信方面,本文中采用的方法是选择两个视角分类器标记一致的子序列作为新一轮的训练数据.这种置信度衡量的一个优点是:随着迭代,分类器输出的噪声将变小,因为标记一致而被选择的数据随之增加,当算法进行到一定程度时,新选择的数据就越来越接近对整个序列进行选择的结果.为了方便比较,我们以直接选择整个句子的协同训练方法作为基线系统.

实验结果见表 4,基于子序列选择的方法在所有的方法中取得了最好的结果(81.52%),相比于直接使用跨语言映射,其结果增长了约 2%,相比于直接应用协同训练,增长了 0.4%(p<0.0001),说明基于片段的数据选择方法能够保证新选择数据具有较高的有效训练数据规模.同时,协同训练方法自身可以减小跨语言映射的噪声.这一点可以从协同训练基线系统和跨语言映射系统的性能比较中观察得到,因为两者在训练最终目标语言模型时使用等量的数据,而协同训练基线系统具有更高的准确率.

Table 4 Comparison of accuracies of different cross-lingual projection systems

表 4 不同跨语言映射系统的性能比较

方法	目标语言词性标注精度(%)	因子图噪声(%)	有效训练数据规模(最终一轮迭代)
只使用目标语言标注数据	65.18	0	–
跨语言映射	79.55	29.55	2 678 869
协同训练(基线系统)	81.14	31.63	3 296 794
基于子序列的协同训练	81.52	27.30	3 508 100

更进一步地,可以基于本文的理论对跨语言映射任务的协同训练方法中样本选择的有效性进行分析.基于第 4.4 节中使用的 CTB 双语语料,我们对不同方法最终训练集上的因子图噪声和有效数据规模进行估计,估计结果同样在表 4 中列出.基于子序列选择的方法具有最低的因子图噪声(27.30%).

5 结论及后续工作

结构化学习由于其标注集合的特殊性,难以被已有的噪声学习理论所涵盖.然而,结构化学习对于噪声可学习性的分析有着更多的需求.本文基于对结构化学习中噪声放大问题的观察,对结构化学习问题的噪声可学习性进行了理论分析,得到了结构化学习问题的噪声可学习性条件,并通过实验验证了理论的正确性.同时,本文对高阶结构化学习模型噪声回退现象的分析以及在跨语言映射和协同训练方法上进行的实验都证明了该理论在现实应用中具有重要的指导意义.

在未来的工作中,我们将主要对以下 3 个方面进行研究:

- 首先,我们将研究如何对本文的理论加以更广泛的推广,使其能够处理判别式结构化学习模型;
- 其次,虽然文献[1]中的证明并没有显示假设数据标记的平衡性,但在具体应用中,数据的类别可能会有极度的不平衡,现有的机器学习理论尚不能描述这种数据的不平衡性会对噪声可学习性产生何种

影响;

- 第三,现实应用中的噪声并非总能满足随机噪声假设,比如在跨领域移植任务中,领域外数据对于领域内模型来说便属于一种噪声数据,但这种噪声与数据类别条件相关,因此是一种固有噪声.在后续的工作中,我们将研究固有噪声环境下的相关理论和算法.

此外,我们将尝试将本文的理论和算法应用到更复杂结构的学习问题中,比如自然语言处理中具有树形输出结构的句法分析任务.

References:

[1] Angluin D, Laird PD. Learning from noisy examples. Machine Learning, 1988,2(4):343−370. [doi: 10.1023/A:1022873112823]

[2] Laird PD. Learning from Good and Bad Data. Boston: Kluwer Academic Publishers, 1988.

[3] Kschischang FR, Frey BJ, Loeliger H. Factor graphs and the sum-product algorithm. IEEE Trans. on Information Theory, 2001, 47(2):498−519. [doi: 10.1109/18.910572]

[4] Yarowsky D, Ngai G. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In: Proc. of the NAACL. 2001. 200−207. http://dl.acm.org/citation.cfm?doid=1072133.1072187

[5] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proc. of the Workshop on Computational Learning Theory (COLT). 1998. [doi: 10.1145/279943.279962]

[6] Kearns M. Efficient noise-tol erant learning from statistical queries. Journal of the ACM (JACM), 1998,45(6):983−1006. http://dl.acm.org/citation.cfm?id=293351

[7] Blum A, Frieze A, Kannan R, Vempala S. A polynomial-time algorithm for learning noisy linear threshold functions. In: Proc. of the 37th Annual IEEE Symp. on Foundations of Computer Science. Burlington, 1996. 330−338. [doi: 10.1109/SFCS.1996.548492]

[8] Cohen E. Learning noisy perceptrons by a perceptron in polynomial time. In: Proc. of the 38th Annual Symp. on Foundations of Computer Science. Miami Beach, 1997. 514−523. [doi: 10.1109/SFCS.1997.646140]

[9] Osborne M. Shallow parsing using noisy and non-stationary training material. Journal of Machine Learning Research, 2002,2: 695−719.

[10] Beigman E, Klebanov BB. Learning with annotation noise. In: Proc. of the Association of Computational Linguistics. Suntec, 2009. 280−287. http://dl.acm.org/citation.cfm?id=1687919

[11] Reidsma D, Akker R. Exploiting ‘subjective’ annotations. In: Proc. of the COLING 2008 Workshop on Human Judgments in Computational Linguistics. Manchester, 2008. 8−16. http://dl.acm.org/citation.cfm?id=1611631

[12] Collins M. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In: Proc. of the New Developments in Parsing Technology. Springer-Verlag, 2005. 19−55.

[13] Taskar B, Guestrin C, Koller D. Max-Margin Markov networks. In: Thrun S, Saul L, Schölkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.129.8439

[14] McAllester D. Generalization bounds and consistency for structured labeling. In: Baklr G, Hofmann T, Schölkopf B, Smola AJ, Taskar B, Vishwanathan SVN, eds. Predicting Structured Data. Cambridge: MIT Press, 2007. 1−16.

[15] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the Int’l Conf. on Machine Learning (ICML). San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[16] Vapnik V. Statistical Learning Theory. New York: Wiley, 1998. 160−196.

[17] Halácsy P, Kornai A, Oravecz C. HunPos—An open source trigram tagger. In: Proc. of the Association of Computational Linguistics. Prague, 2007. 209−212. http://dl.acm.org/citation.cfm?id=1557830

[18] Xue N, Xia F, Chiou FD, Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural Language Engineering, 2005,11(2):207−238. [doi: 10.1017/S135132490400364X]

[19] Hu P, Yu M, Li J, Zhu C, Zhao T. Semi-Supervised learning framework for cross-lingual projection. In: Proc. of the Web Intelligence/IAT Workshops 2011. 2011. 213−216. [doi: 10.1109/WI-IAT.2011.58]

[20] Toutanova K, Klein D, Manning C, Singer Y. Feature-Rich part-of-speech tagging with a cyclic dependency network. In: Proc. of the HLT-NAACL 2003. 2003. 173−180. http://dl.acm.org/citation.cfm?id=1073478.

[21] Och FJ, Tillmann C, Ney H. Improved alignment model. In: Proc. of the 38th Annual Meeting on Association for Computational Linguistics. 2000. 440–447. http://dl.acm.org/citation.cfm?id=1075274

[22] Petrov S, Das D, McDonald R. A universal part-of-speech tagset. In: Proc. of the 8th Int'l Conf. on Language Resources and Evaluation (LREC) 2012. 2012. http://www.ryanmcd.com/papers/uposLREC2012.pdf

附录 1:定理 2 的证明过程

对于定理 1,文献[1]所给出的证明(见文献[1]的定理 2)分为两个步骤:

(1) 计算得到每一个假设在训练集上的输出与样本标记不一致的概率;

(2) 基于上述概率计算所有与真实函数 L_*差异大于ε的假设(ε-bad 假设,即 $d(L_i,L_*)\geqslant\varepsilon$的所有假设 L_i)在训练过程中被选中的概率.

上述过程的第 2 步只考虑假设的输出与样本标记是否一致,被看作是一个二项分布问题,因此与学习问题的类别数无关.基于此,为了证明定理 2,我们只需得到 C 类分类问题中各假设与样本标记不一致的概率,将这一概率带入到第 2 步的证明过程中,即可得到结论.

证明:对于假设空间中的任意假设 L_i,无论训练集是否有噪声,其与最优假设 L_*不一致的概率均为

$$d_i=d(L_i,L_*).$$

当样本存在噪声 η时,对于任何一个给定样本,L_i 与样本标记之间不一致可以分为两种情况:

- 首先,在 L_i 与 L_*输出一致的情况下,若有 L_i 与样本标记之间不一致,需满足样本标记有噪声,这一事件对应的概率为$(1-d_i)\eta$;
- 当 L_i 与 L_*输出不一致时,这一情况可以进一步细分为两类:
  (1) 样本标记正确,对应的概率为 $d_i(1-\eta)$,或者
  (2) 样本标记错误,错误标记为除 $L_i(x)$和 $L_*(x)$以外的 $C-2$ 个类别中的任意一个,对应的概率为
  $$d_i\eta\times(C-2)/(C-1).$$
  从而,L_i 与样本标记之间不一致的概率为

$$\begin{aligned}p_i &= (1-d_i)\eta + d_i(1-\eta) + d_i\eta\times(C-2)/(C-1)\\ &= \eta + d_i\left(1-\frac{C}{C-1}\eta\right)\\ &\geqslant \eta + \varepsilon\left(1-\frac{C}{C-1}\eta\right).\end{aligned}$$

之后的证明与文献[1]中的定理证明相同. □

于墨(1986－),男,黑龙江黑河人,博士生,CCF 学生会员,主要研究领域为机器学习,句法分析.
E-mail: yumo@mtlab.hit.edu.cn

胡鹏龙(1988－),男,硕士生,主要研究领域为机器学习.
E-mail: penglonghu@gmail.com

赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器翻译,句法分析,机器学习.
E-mail: tjzhao@hit.edu.cn

郑德权(1968－),男,博士,副教授,CCF 会员,主要研究领域为自然语言处理,信息检索,机器学习.
E-mail: dqzheng@mtlab.hit.edu.cn