计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131113
52(1):191-199, 2015

基于深度学习框架的隐藏主题变量图模型

吴 蕾[1] 张文生[2] 王 珏[2]
[1](中国农业科学院农业信息研究所 北京 100081)
[2](中国科学院自动化研究所 北京 100190)
(girlrable@126.com)

Hidden Topic Variable Graphical Model Based on Deep Learning Framework

Wu Lei[1], Zhang Wensheng[2], and Wang Jue[2]
[1](*Agricultural Information Institute*, *Chinese Academy of Agricultural Sciences*, *Beijing* 100081)
[2](*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract The hidden topic variable graphical model represents potential topics or potential topic changes by nodes. The current study of hidden topic variable graphical models suffers from the flaw that they can only extract single level topic nodes. This paper proposes a probabilistic graphical model based on the framework of deep learning to extract multi-level topic nodes. The model adds the preprocessing layer to the bottom of the hidden topic variable graphical model. The preprocessing layer used in the paper is the self-organizing maps (SOM) model. By introducing the SOM, the model can effectively extract different topic status with those extracted by the hidden topic variable graphical model. In addition, the hidden topic variable graphical model used in this paper is constructed by hidden Markov model (HMM) and conditional random field (CRF). In order to make up the short-distance dependency Markov property, we use the characteristic function defined by first-order logic. On this basis, we propose a new algorithm by hierarchically extracting topic status. Experimental results on both the international universal Amazon sentiment analysis dataset and the Tripadvisor sentiment analysis dataset show that the proposed algorithm improves the accuracy of sentiment analysis. And the new algorithm can mine more macroscopic topic distribution information and local topic information.

Key words probabilistic graphical model; deep learning; hidden Markov models (HMM); self-organizing maps; first-order logic

摘 要 隐藏主题变量图模型是一种用节点表示潜在主题或者潜在主题变化的概率图模型.针对当前隐藏主题变量图模型只能提取单层主题节点的缺陷,提出一种基于深度学习框架的提取多层主题节点的概率图模型.该模型在隐藏主题变量图模型的底层增加预处理结构层,即引入自组织映射层,可以有效地提取不同层次的主题状态.另外,隐藏主题变量图模型使用了隐马尔可夫网络和条件随机场的相结合的模型.针对条件随机场,提出了一阶逻辑子句定义的特征函数.弥补了长距离依存特性的缺失.在此基础上提出了一种分层次提取主题状态的新深度学习算法.在国际通用的亚马逊情感分析数据、Tripadvisor情感分析数据上的实验表明,新算法可以提升情感分析的准确率.同时实验结果也表明,提取多层主题状态可以更好地挖掘宏观主题分布信息和评论的局部主题信息.

收稿日期:2013-07-30;**修回日期**:2014-03-28
基金项目:国家自然科学基金重点项目(U1135005);国家自然科学基金重大研究计划项目(90924026);国家自然科学基金青年科学基金项目(61305018);国家科技重大专项项目(GFZX0101050302);武器装备预研基金项目(51301010206)

关键词 概率图模型;深度学习;隐马尔可夫模型;自组织映射;一阶逻辑

中图法分类号 TP181; TP183

随着诸如博客、微博、社交网络等社会媒体的出现,近年来由网络用户产生的评论、评价等级以及推荐数据大量涌现.例如,来自 Amazon.com 的用户对购买商品的评价以及来自 TripAdvisor.com 的关于出行、住宿的评价.利用好这些数据可以使商业界更好地出售商品或者服务,甚至发现新的市场机遇.为了更好地利用这些数据,产生了一个新的研究领域——情感分析.然而概率图模型的方法正可以处理这类问题.

隐藏变量是已知它的存在而没有机会直接观测的数据变量.在概率图模型中隐藏变量发挥了巨大的作用.一方面它们可以作为父节点解释观测变量;另一方面它们可以简化网络学习的复杂性[1].因此在情感分析中引入表示主题的隐藏变量(即判断用户对商品或者服务的情感态度属于哪个给定的主题[2])来学习标注或者打分信息已经成为概率图模型来处理此类问题的趋势.但是当前的隐藏主题变量概率图模型只提取单层次主题变量.单层次主题变量往往只能从 1 个层面上揭示观测数据的成因.

近年来深度学习作为一种新的研究领域越来越受到研究者们的关注.深度学习源于神经网络,它是一种含有多层隐含层的多层感知器.同时深层结构可以被看作是具有多层隐藏变量节点的概率图模型[3].但是在当前的深度学习模型中,不同层之间的节点往往是全连接,这就增加了计算复杂度,并降低了模型的泛化能力.

针对以上 2 点本文提出一种可以提取多层次隐藏主题变量的概率图模型.该模型具有 2 层结构:第 1 层是自组织映射层,目的是提取获胜节点的权重,这里称之为初级主题状态;第 2 层将每个单词对应的初级主题状态作为输入,输入到隐马尔可夫网和条件随机场的联合网络.通过该层网络可以进一步提取高级主题状态节点,并通过 2 种不同层次的主题状态获得标签信息.另外,本文提出的模型中输入层的单词节点与初级主题状态节点是 1 对 1 的连接,初级主题状态节点与高级主题状态节点是 1 对 1 的连接,主题状态节点与标签节点也是 1 对 1 的连接.这就能使网络变得更稀疏.

但是条件随机场的马尔可夫性质导致长距离依存关系缺失问题[4].而且现有的深度学习大多采用词包(bag of words)的形式.这种词包的形式没有考虑词序的信息,并假设所有词都是独立同分布.这是不符合自然语言特点的.针对以上 2 个缺点,本文使用了一阶逻辑子句来定义条件随机场的特征函数,即表示标签与 2 种主题状态之间的关系.一阶逻辑是一种表示知识的简单方式.在一阶逻辑的定义时本文考虑了初级主题状态节点的同现关系(不考虑初级节点之间距离)以及标签的顺序关系.然后通过概率计算可以得到关于标签的后验概率.另外在隐马尔可夫过程中高级主题状态节点之间的状态转移矩阵也存在顺序关系.

1 相关工作

1.1 情感分析

情感分析的任务是在大量文本中发现并理解主观信息.它包含 2 个方面:1)句子标注任务[1,5-10],该任务的目标是为每个句子标注它们讲述的内容标签;2)评级预测任务[6-7,11-13],该任务是为每条评论打分.分数越高这条评论的含义就越正面.以往在这 2 方面的工作大多是文档级别或者句子级别的分类,但是这种方法无法处理文档或者句子中出现多维度标注信息的情况[7].另外研究者们通常人工地向模型加入设计好的负面单词或者具体的词性信息.但是它们的效果都不明显.前人的文章里介绍了一种短语级别的分类[14],并引入了一种设计巧妙的特征:依存树.本文处理的是单词级别的情感分类问题,并加入了一阶逻辑子句作为特征函数,其中子句的基原子可以被看作特征.由于篇幅有限,更多关于情感分类的细节请参考文献[15].

1.2 隐藏主题变量图模型

对于含有隐藏主题变量,隐藏层不足以表示不同层次特征的概率图模型[6-7,14,16-20].这些模型中的主题节点表示每个观测变量的潜在主题.模型假设所有观测变量都是由隐藏主题产生的.文献[16]展示了将局部基于实体的方法加在基于隐马尔可夫的内容模型上的方法.文献[17]描述了一种内容结构贝叶斯模型,其主题需要从相关文本中选择.文献[7,18]的作者们研究了改进隐藏狄利克雷分配(latent dirichlet allocation, LDA)[19]的概率主题模型,它用于提取文本摘要.隐藏变量是由单词出现频率决定的.然而这种使用词包的方法存在上文提到的缺点.

另外在基于条件随机场的隐藏变量结构的研究中,也出现了许多优秀的工作.其中,隐藏单元条件随机场(hidden-unit CRF, HUCRF)[20]是一种监督学习的模型,隐藏单元被设置在观测变量和标签变量之间.然而标注数据需要花费大量人力物力.文献[6]介绍了一种将隐马尔可夫网和条件随机场相结合的联合内容模型(joint content model, JointCM).该模型使用标注和非标注语料,而不需要过多地借助人工设置或者仅仅使用标注语料.但是这种方法的观测节点之间没有线连接.这种稀疏结构使我们可以向模型加入关于观测变量的更加丰富的一阶逻辑表示信息.这样可以帮助理解观测变量复杂的依存关系.

1.3　深层网络

含有多层隐藏节点的深层网络是近年来热门的研究领域.出现了大量研究成果:深层信度网络(deep belief network, DBN)[21]是由多层有向信度网络单元和1层双向受限玻尔兹曼机(restricted Boltzmann machine, RBM)单元构成的.受限玻尔兹曼机可以看作是一种特别的马尔可夫随机场(Markov random field, MRF).单元之间是全连接,没有节点层内部节点的连接.在网络中每个单元的输出作为下一个单元的输入.深层凸网络(deep convex networks, DCN)[22]是由多层模块构成的.每层模块具有3层子结构:第1层子结构负责线性输入特征;第2层子结构是隐藏的非线性单元;第3层子结构负责线性输出原始输入数据和模型输出数据.最后1层模块的第3层子结构负责输出目标数据.堆去噪自动编码(stacked denoising auto-encoders, SDA)[23]是一种用于领域自适应的深层学习方法,它将单层去噪自动编码的输出作为输入堆叠起来的深层结构.

1.4　一阶逻辑

一阶逻辑知识库是一阶逻辑子句或公式构成的集合.作为一种知识表示的方式,一阶逻辑已用于马尔可夫网络.马尔可夫逻辑网[24]就是由一阶逻辑和马尔可夫网络构成的.其中一个一阶逻辑子句是网络中的一个团(clique),子句中每个基原子是一个节点.马尔可夫逻辑网为每个一阶逻辑子句分配一个权重.这就使不满足一阶逻辑子句的事件也有可能存在,只是概率会小些.参照马尔可夫逻辑网,本文将一阶逻辑用于定义条件随机场的特征函数,从而增加了变量之间的依存关系的信息.

2　隐藏主题变量图模型

本文使用了文献[6]中的隐马尔可夫网和条件随机场相结合的图模型.下面首先介绍本文使用的符号.我们用 $\boldsymbol{W}=(\boldsymbol{W}_1,\cdots,\boldsymbol{W}_m)$ 表示一篇文档,$\boldsymbol{W}_j=(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{m'})$ 表示一个句子,其中 $\boldsymbol{w}_i$ 表示文本中的单词.$\boldsymbol{Y}=(y_1,\cdots,y_{m'})$,其中 y_i 表示单词对应的标签.$\boldsymbol{T}=(t_1,\cdots,t_{m'})$ 表示相应的主题序列.主题的种类数事先由人工设定.

单词级别 JointCM 方法用隐马尔可夫模型描述单词和隐藏主题变量之间的关系:隐藏主题变量之间存在状态转移,且每个单词 $\boldsymbol{w}_i$ 由一个相应的隐藏主题变量 t_i 产生.其联合概率为

$$P_\theta(\boldsymbol{W}_j,\boldsymbol{T})=\prod_{i=1}P_\theta(t_i\mid t_{i-1})P_\theta(\boldsymbol{w}_i\mid t_i). \quad (1)$$

然后,由隐马尔可夫预测的主题和单词节点被共同放入条件随机场中,并构建它们与标签的关系.隐藏主题节点和单词节共同点与标签节点相连接,标签节点的后验概率为

$$P_\phi(\boldsymbol{Y}\mid\boldsymbol{W}_j,\boldsymbol{T})=\prod_{i=1}P_\phi(y_i\mid\boldsymbol{w}_i,t_i). \quad (2)$$

最后,将2个模型通过节点相连,且联合概率由隐马尔可夫模型的概率和条件随机场的概率构成,其联合概率为

$$P(\boldsymbol{W}_j,\boldsymbol{T},\boldsymbol{Y})=P_\theta(\boldsymbol{W}_j,\boldsymbol{T})P_\phi(\boldsymbol{Y}\mid\boldsymbol{W}_j,\boldsymbol{T}). \quad (3)$$

3　基于深度学习的图模型算法

图1是改进隐藏主题变量图模型的深层网络图,观测层是输入的单词.第1层隐藏层节点表示初级主题状态节点,第2层隐藏层节点表示高级主题

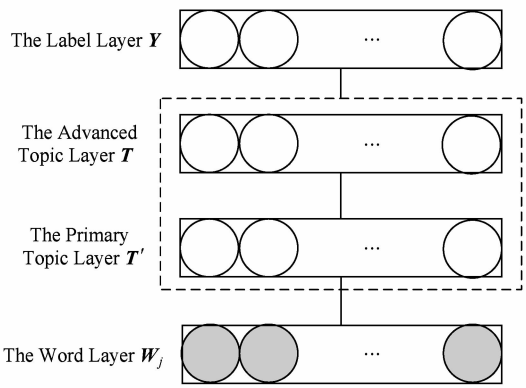

Fig. 1　The deep network of hidden topic variables graphical model.

图1　基于深度学习的隐藏主题变量图模型

状态节点.这些状态可以看作是对单词不同层次的表示.每个输入单词对应1个初级主题节点和1个高级主题节点,有可能对应多个标签节点.但本文中算法只为每个单词选取后验概率最大的标签.本文从学习和推理2部分对改进模型进行了讨论.

3.1 改进模型的学习算法

这里我们将分层讨论改进模型的学习问题.首先,将已知单词节点层输出的单词信息输入到初级主题节点层,初级主题层的输出即为初级主题信息.在这里初级主题层利用自组织映射(self-organizing maps, SOM)完成从单词到初级主题的提取.自组织映射是一种将高维数据流形非线性,有序并且光滑地映射到具有固定结构的低维阵列的方法[25-26](本文以二维为例).本文使用的自组织映射采用随机梯度下降方法,具体过程如下:

1) 令某输入单词为 $\boldsymbol{w}\in\mathbb{R}^n$,表示单词有 n 维特征.对于低维陈列中的每个节点给出一个权值 $\boldsymbol{t}'_i\in\mathbb{R}^n$.求出与单词距离最小的权值及其对应节点,即 $c=\arg\min_i d(\boldsymbol{w},\boldsymbol{t}'_i)$.

2) 根据梯度下降更新每个节点的权值为

$$\boldsymbol{t}'_i(k+1)=\boldsymbol{t}'_i(k)+h_{ci}(k)(\boldsymbol{w}-\boldsymbol{t}'_i(k)),$$

$$h_{ci}(k)=\alpha(k)\exp\left(-\frac{\|\boldsymbol{r}_c-\boldsymbol{r}_i\|^2}{2\sigma^2(k)}\right),$$

其中,$\boldsymbol{r}_c,\boldsymbol{r}_i\in\mathbb{R}^2$ 分别表示二维阵列上2个节点的位置.自组织映射提供了几种低维阵列的拓扑结构,例如六边形阵列、矩形阵列.低维阵列的拓扑结构一旦确定距离便可求出.

3) 对新来的单词,更新参数 $\alpha(k)$,重复上述步骤1).

本文中令 $\alpha(k)=\alpha\exp(-\beta k/k_{\max})$.该参数控制了获胜节点对邻域节点的影响强度,强度随着迭代次数增加而快速减小,并在降到较小的正值时缓慢趋于0.另外当满足 $\sum h_{ci}(k)=+\infty$ 且 $\sum h_{ci}(k)^2<+\infty$ 时,迭代中 $\boldsymbol{t}'_i$ 的增加和减少冲突的次数有限,所以 $\boldsymbol{t}'_i\to\boldsymbol{t}'^*$ 算法收敛[25].自组织映射过程的复杂度为 $O(Ik_{\max})$,其中 I 为节点个数,$k_{\max}$ 为循环最大次数.当循环结束就得到一系列节点的权重.这些权重就是初级主题状态.将每个单词对应节点的初级主题状态作为下层的输入,开始第2层的学习.

第2层将第1层得到的初级主题状态 $\boldsymbol{T}'$ 作为输入,将高级主题状态 $\boldsymbol{T}$ 作为隐藏层,标签 $\boldsymbol{Y}$ 作为输出层.这里我们使用JointCM方法来处理.第1层并没有用到标签信息,所以对于标注和非标注文本同样看待.然而第2层里使用了标注文本的标签信息.对于标注文本 D_{L},对数似然为

$$L_{\mathrm{L}}(\theta,\omega)=\sum_{(\boldsymbol{T}',\boldsymbol{Y})\in D_{\mathrm{L}}}\ln P(\boldsymbol{Y},\boldsymbol{T}')=\sum_{(\boldsymbol{T}',\boldsymbol{Y})\in D_{\mathrm{L}}}\ln\sum_{\boldsymbol{T}}P(\boldsymbol{Y},\boldsymbol{T},\boldsymbol{T}'),\tag{4}$$

其中,$\boldsymbol{T}'$ 表示初级主题状态,$\boldsymbol{T}$ 表示高级主题状态,$\boldsymbol{Y}$ 表示标签.我们使用EM算法优化目标,过程如下.

1) E-步骤:计算隐藏变量 $\boldsymbol{T}$ 的后验分布 $P(\boldsymbol{T}|\boldsymbol{Y},\boldsymbol{T}',\theta)$,其中:

$$P(\boldsymbol{T}',\boldsymbol{T},\boldsymbol{Y})=P_\theta(\boldsymbol{T}',\boldsymbol{T})P(\boldsymbol{Y}|\boldsymbol{T}',\boldsymbol{T})=\left(\prod_{i=1}P_\theta(t_i|t_{i-1})P_\theta(\boldsymbol{t}'_i|t_i)\right)\left(\prod_{i=1}P(y_i|t_i,\boldsymbol{t}'_i)\right).\tag{5}$$

如果没有 $P(y_i|t_i,\boldsymbol{t}'_i)$,式(5)可以看作隐马尔可夫过程,并用前向-后向算法解决这个问题.显然在隐马尔可夫模型中隐藏状态(即本文中的高级主题状态)是存在转移概率的,因此在本模型中存在高级节点的顺序关系.我们单独分析了 $P(y_i|t_i,\boldsymbol{t}'_i)$.这里引入了3种类型的一阶逻辑子句的基原子:主题与标签、标签之间以及初级主题之间的基原子.因此,条件分布可以表示成如下形式:

$$P(\boldsymbol{Y}|\boldsymbol{T}',\boldsymbol{T})=\frac{1}{Z}\prod_{c\in C}\Phi_c(y,t,t'),$$

这里,Z 是标准化常数,C 是所有因子的范围.改进的隐藏内容结构模型需要将一阶逻辑加入条件随机场.为了将基于隐藏内容结构的文本分析映射为条件逻辑,改进方法定义了如下3种子句:

$$Pr_{\mathrm{label}}(t,y)\wedge Pr_{\mathrm{label}}(\boldsymbol{t}',y)^+;$$

$$Pr_{\mathrm{labelnext}}(y_{-1},y)^+;$$

$$\neg Pr_{\mathrm{same}}(\boldsymbol{t}',\boldsymbol{t}'_+)\vee\neg Pr_{\mathrm{label}}(\boldsymbol{t}'_+,y)\vee Pr_{\mathrm{label}}(\boldsymbol{t}',y)^*.$$

用+标注的子句适用于短距离依存特点;第3个用*标注的子句适用于长距离依存特点;$t,\boldsymbol{t}'$ 和 y 分别表示初级主题、高级主题以及标签;$\boldsymbol{t}'_+$ 表示不同于 $\boldsymbol{t}'$ 的初级主题;y_{-1} 表示前1个状态所对应的标签;析取符号 $\vee$ 表示子句中至少有1个基原子为真时子句为真;合取符号 $\wedge$ 表示当且仅当子句中每个基原子都为真,那么子句才为真;取反符号 $\neg$ 表示对基原子取反;第1个子句表示初级主题状态 $\boldsymbol{t}'$ 和高级主题状态 t 的标签是 y;第2个子句表示标签 y 对应主题状态的前1个主题状态的标签是 y_{-1};第3个子句表示如果初级主题状态 $\boldsymbol{t}'$ 和 $\boldsymbol{t}'_+$ 相同,并且 $\boldsymbol{t}'_+$ 的标签是 y,那么 $\boldsymbol{t}'$ 的标签也是 y.这里不限定 $\boldsymbol{t}'$ 和 $\boldsymbol{t}'_+$ 必须连续出现,因此可以弥补长距离依存信息的缺失.上面的条件分布就可以表示成式(6):

$$P(y|t,\boldsymbol{t}')=\frac{1}{Z}\exp\Big(\sum_i \omega_i n_i(y,t,\boldsymbol{t}')\Big),\quad(6)$$

这里,Z 是标准化常数,ω_i 是第 i 个子句的权重,$n_i(y,t,\boldsymbol{t}')$表示满足条件的真子句个数.

通过使用技巧[27]可以得到 3 种子句的权重分别为 $\ln(n_{y,t,t'}/n_{t,t'})$,$\ln(n_{y_{-1},y}/n_{y_{-1}})$以及 $\ln(n_{t',y,t'_+,y}/n_{t'_+,y})$,其中 $n_{t,t'}$表示初级主题 $\boldsymbol{t}'$和高级主题 t 同现的次数;$n_{y,t,t'}$ 表示满足第 1 个子句的状态个数;$n_{y_{-1},y}$表示满足第 2 个子句的状态个数;$n_{y_{-1}}$ 表示标签 y_{-1}出现的次数;$n_{t',y,t'_+,y}$和 $n_{t'_+,y}$分别表示满足第 3 个子句的状态个数以及满足第 3 个子句的假设部分的状态个数.根据以上定义可以得到对于每个公式都有 $\sum \exp(\omega_i)=1$,其中 ω_i 是上面定义的权重.使用这个技巧算法可以避免陷入局部极小值.

2) M-步骤:由于前文中已经固定了参数 ω 的值,这里不再考虑.因此,这里仅仅需要确定主题参数 θ.在 E-步骤中发散分布和转移分布均已被更新.至此,所有问题都可解决.

考虑到存在大量非标注语料,我们加入了非标注语料的信息.标注和非标注文本的目标函数如下式所示:

$$L(\theta)=L_{\mathrm{U}}(\theta)+L_{\mathrm{L}}(\theta).$$

对于非标注文本 D_{U},对数似然为

$$L_{\mathrm{U}}(\theta)=\sum_{\boldsymbol{T}\in D_{\mathrm{U}}}\ln P_\theta(\boldsymbol{T}')=\sum_{\boldsymbol{T}'\in D_{\mathrm{U}}}\ln\sum_{\boldsymbol{T}}P_\theta(\boldsymbol{T}',\boldsymbol{T}).$$

没有标签的内容结构模型就是一个隐马尔可夫过程,使用前向-后向方法就可以解决.

序列(θ^t,ω^t)令对数似然函数 $L(\theta^t,\omega^t)$的取值不断增加,直到收敛到稳定值.具体的证明可以参考文献[28],这里只给出简要证明思路:因为 $L(\theta^{t+1},\omega^{t+1})\geqslant L(\theta^t,\omega^t)$,且采样个数有限,因此增长序列 $L(\theta^t,\omega^t)$只有有限数值,即会收敛到稳定值.第 2 层在最坏情况下的复杂度为 $O(M^2T+|T|^2Y+Y^2)$,其中 M 为隐藏状态种类数,T 为初级主题状态种类数,$|T|$为初级主题状态节点个数,Y 为标签种类数.

传统的深度学习不考虑同层节点之间的连接关系,或者说只考虑的是前层节点与下层节点之间的量化关系.本文通过使用图模型的方法在计算隐马尔可夫过程时需要考虑高级主题状态节点之间的前后关系信息即转移概率,又在定义一阶逻辑子句中考虑了初级主题状态的同现关系信息,即长距离依存信息.因此图模型的优点恰好可以弥补传统深度学习同层节点之间信息损失的问题.

3.2　改进模型的推理算法

在第 1 层中:对于每个新来的单词,寻找与其距离最小的获胜节点的权重,并作为此新单词的初级主题节点.

在第 2 层中:在给定初级主题节点以及参数的情况下,运用 Viterbi 算法可以计算高级主题状态.再根据 2 种主题状态,同样用 Viterbi 算法可以得到新单词的标签.

4　实验与分析

4.1　实验设置

为验证本文方法的有效性,我们分析了 2 种分类问题:训练集与测试集分布相同的简单分类问题,以及训练集与测试集分布不同的领域自适应分类问题[29].

对于简单的分类问题,我们将本文的方法和其他 3 种方法进行了比较.3 种方法分别是 LDA,HUCRF 以及 JointCM.实验的目标是解决 2 种情感分析任务:句子标注任务及评级预测任务.实验分别采用的 2 个数据集如下:

1) 亚马逊数据[30]:该数据集包括用户对于 22 个不同产品类型的 34 万条评论.标签是{正,负}2 种.我们选取其中 4 个不同领域:{Books,DVDs,Electronics 和 Kitchen}.每个领域的评论包括 1 000 条正评论、1 000 条负评论以及 3 000～5 000 条非标注评论.对每个领域我们在标注语料里随机选取 1 200 条评论作为标注训练集,800 条作为测试集,1 000 条作为非标注训练集.正、负评论条数比皆为 1∶1.

2) TripAdvisor 数据[13]:该数据集给出了每条评论的总体打分以及对每条评论从 7 个方面(即 value, room, location, cleanliness, check-in/front desk,service 和 business services)分别打分.分数是从 1～5 的整数打分(例如,1～5 颗星).1 颗星表示很不满意,5 颗星表示很满意.实验中我们按照文献[13]的方法去掉了过短以及过长的评论以及打分不全的评论,最后得到 66 512 条评论.我们随机选取其中的 50%作为标注训练集,20%作为测试集,剩下的作为非标注训练集.

实验中 JointCM 方法以及本文提出的改进方法都使用了所有训练集合进行训练.LDA 算法使用了所有训练文本聚类,然后用投票的方式只计算测试集上的结果.HUCRF 只使用标注训练集进行训练.

领域自适应分类问题是一种源领域和目标领域分布不同的迁移学习.解决方法是假设 2 个领域

具有相同的中间概念(即隐藏层).例如商品价格、商品质量、卖家的服务等.这里使用领域自适应分类问题来验证算法,是因为情感分析的领域自适应问题处理的好坏能够说明深层学习到底在多大程度上解决了数据表示的复杂性.我们将本文的方法与其他3种方法进行了比较,分别为结构一致学习(structural correspondence learning, SCL)[30], SDA以及JointCM方法.SCL是一种有效的针对自适应问题的方法.该方法通过同现频率或者互信息等测量方式寻找2个领域的共有关键特征,再以此特征进行学习.SDA方法是一种标准的深层学习算法.该方法一旦参数确定了就可以通过梯度下降方法初始一个有监督的神经网络,或者将高层特征输入给一个分类器进行计算.本次实验仍然采用亚马逊数据集.

4.2 实验结果与分析

本文第1个数据集使用的评价指标是准确率(accuracy).准确率=正确分类个数/事例总数.第2个数据集使用的评价指标是L1误差[6]以及ρ_{aspect}, ρ_{review}和MAP@10[13].其中L1误差指模型预测和真实值之间差的绝对值,这里我们已经将结果归一化到0~1之间,L1误差越小表示结果性能越好.ρ_{aspect}和ρ_{review}是2个模型预测和真实值之间的平均皮尔逊相关.前者描述了对于某条评论,模型对各个方面的预测能力(例如,评论者是否更看重位置而不太重视清洁度),后者描述了在某个方面模型对于每条评论的预测能力(例如,各个酒店在食物方面的评价情况).其中2个指标定义如下:

$$\rho_{\text{aspect}}=\frac{\sum_{d=1}^{|D|}\rho_{\boldsymbol{s}_d,\boldsymbol{s}_d*}}{|D|},$$

$$\rho_{\text{review}}=\frac{\sum_{k=1}^{K}\rho(\boldsymbol{S}_k,\boldsymbol{S}_k*)}{K},$$

其中,$|D|$表示样本总数,K表示每条评论是从几个方面给出打分的,这里$K=8$. MAP@10表示在真实评论中分数最高的10条评论,被模型预测后依然在Top10的比率.

$$\text{MAP@10}=\frac{\sum_{k=1}^{K}(\text{仍然在 Top10 的评论个数}/10)}{K}.$$

对于简单分类问题,图2给出了4种方法在准确率上的比较.可见通过提取多层次主题状态再进行分类学习,并利用一阶逻辑信息可以提高原分类方法的性能,甚至对于原分类方法不能较好处理的数据(例如,图2所示的DVDs数据)也能给出较好的结果.其中HUCRF的准确率在4种任务上都不如其他方法,这是由于该方法训练时只使用标注训练集,而缺少大量非标注语料的信息.

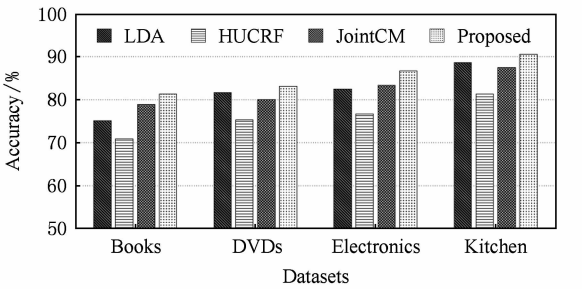

Fig. 2 Accuracy of the sentence labeling task on the Amazon corpus.

图2 亚马逊语料上句子标注任务的准确率

表1给出了4种方法在评价预测任务上的结果.由2种皮尔逊参数的结果可见4种方法对宏观的评价预测性能要好于分项预测性能,即能够较好地预测每条评论的喜好趋势,但是对具体的某条评论的多个方面就不能给出较好的预测结果.黑体字标出了每种评价指标下最好的2种结果.我们的方法在每种指标下都取得了较好的结果.另外LDA在L1误差和ρ_{review}指标下也给出了较好的结果.这表明LDA比较擅长预测评论的总体趋势.HUCRF方法更擅长推荐前几个评价最高的酒店,而这项指标显然对用户更有用.JointCM比较擅长区分每条评论的各个方面.这是由于LDA是从宏观的概率上模拟每条评论的主题分布和评论中每个单词的主题分布,所以可以给出较好的整体趋势预测.JointCM方法中由于使用到隐马尔可夫网络的隐藏主题状态转移矩阵,所以对在已知当前状态下调整下条状态很有帮助,因此在面对同一评论的不同局部打分时,此方法可以较好地作出调整.而通过加入分层次提取的主题状态以及一阶逻辑子句强表示能力,本文提出的方法可以使JointCM在各个方面的性能都有提高,这表明本文的方法可以从整体和

Table 1 The Rating Prediction Results for Tripadvisor Data

表1 Tripadvisor数据上的评级预测结果

Methods	L1 Error	ρ_{aspect}	ρ_{review}	MAP@10
LDA	**0.129**	−0.149	**0.454**	0.143
HUCRF	0.471	0.012	0.208	**0.227**
JointCM	0.133	**0.213**	0.432	0.211
Proposed	**0.115**	**0.413**	**0.637**	**0.357**

局部等多方面挖掘主题信息.

对于领域自适应分类问题,图 3～6 给出 4 种方法在亚马逊语料上的迁移准确率:

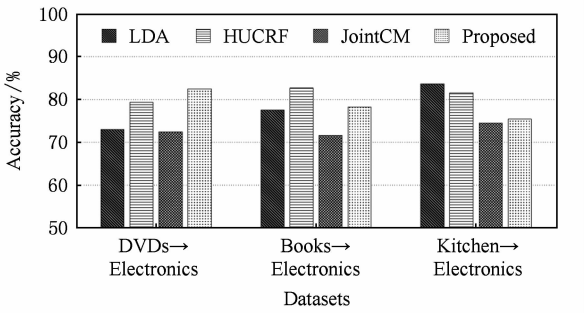

Fig. 3　Accuracy for domain adaptation using the Electronics corpus as the test set.

图 3　Electronics 语料测试集领域自适应问题准确率

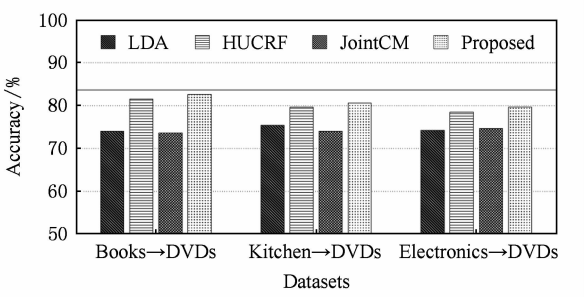

Fig. 4　Accuracy for domain adaptaion using the DVDs corpus as the test set.

图 4　DVDs 语料测试集领域自适应问题准确率

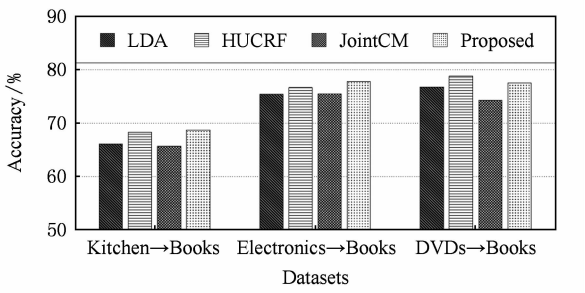

Fig. 5　Accuracy for domain adaptation using the Books corpus as the test set.

图 5　Books 语料测试集领域自适应问题准确率

在图 3～6 中,"M→N"表示用数据集 M 的训练集和非标注集训练,然后在数据集 N 的测试集上测试. 图 3～6 中直线给出了使用本文提出的方法在目标领域训练及测试的准确率. 可见只有在 Electronics→Kitchen 任务上迁移的准确率要最接近(但仍低于)非迁移准确率,在其他任务中迁移准确率都仍远远低于本文方法的非迁移准确率. 一方面这是由于本文方法在该数据集的非迁移问题上的准确率已经

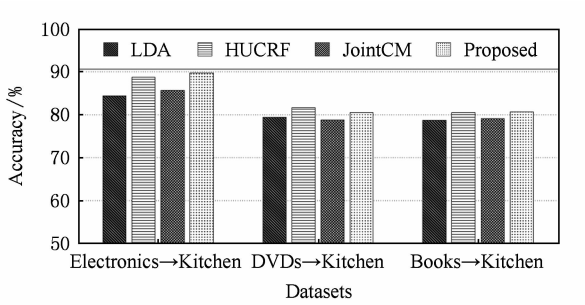

Fig. 6　Accuracy for domain adaptation using the Kitchen corpus as the test set.

图 6　Kitchen 语料测试集领域自适应问题准确率

比其他几种对比方法高;另一方面是因为迁移问题本身由于训练和测试语料的差异性导致算法的性能会比在非迁移问题上的性能低. 这也是下一步研究的方向:在保证非迁移准确率的情况下,逐步提高迁移准确率. 另外由实验可见本文提出的方法在大多数迁移任务中的迁移准确率都高于其他方法. 并且在 11 个任务上本文的方法和 SDA 方法的性能要高于 SCL 方法和 JointCM 方法的性能. 这表明加入基于分层结构提取多层次主题状态或特征的方法可以挖掘出更多更有用处的关于观测变量和标签变量之间的主题信息.

5　总　　结

本文提出了一种基于深度学习框架,但每层使用的算法不同于已有深度学习算法的模型. 本文提出的模型首先使用自组织映射网络非监督地提取初级主题状态,然后将这些初级主题状态输入到经过一阶逻辑改进的 JointCM 模型中. 在 JointCM 模型中再次提取了高级主题状态. 基于深度学习的改进隐藏内容结构模型通过分层次提取主题状态,对深度模型的每层都赋予相对应的物理意义,加强了深度学习的可解释性;通过加入同层次主题状态之间的顺序以及同现关系信息,弥补了深度模型同层节点无序所造成的信息缺失问题. 实验将模型应用到情感分析中,取得了良好的结果. 今后的工作将继续针对自适应领域进行展开,以求得在保证非迁移准确率的前提下继续提高迁移准确率.

参　考　文　献

[1] Kollar D, Friedman N. Probabilistic Graphical Models Principles and Techniques [M]. Cambridge, Massachusetts: The MIT Press, 2009: 713-715

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques [C] //Proc of the ACL-02 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2002: 79-86

[3] Bengio Y. Learning deep architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127

[4] Wu Z J, Yu Z T, Guo J Y, et al. Fusion of long distance dependency features for chinese named entity recognition based on Markov logic networks [G] //Natural Language Processing and Chinese Computing. Berlin: Springer, 2012: 132-142

[5] Pang B, Lee L. Seeing stars: Exploring class relationships for sentiment categorization with respect to rating scales [C] //Proc of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: ACL, 2005: 115-124

[6] Sauper C, Haghighi A, Barzilay R. Incorporating content structure into text analysis applications [C] //Proc of the 2010 Conf on Empirical Method in Natural Language Processing. Stroudsburg, PA: ACL, 2010: 377-387

[7] Lu B, Ott M, Cardie C, et al. Multi-aspect sentiment analysis with topic models [C] //Proc of the 11th IEEE Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2011: 81-88

[8] McDonald R, Hannan K, Neylon T, et al. Structure models for fine-to-coarse sentiment analysis [C] //Proc of the 45th Annual Meeting of the Association of Computational Linguistics. Stroudsburg, PA: ACL, 2007: 432-439

[9] Sauper C, Haghighi A, Barzilay R. Content models with attitude [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2011: 350-358

[10] Dave K, Lawrence S, Pennock D M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [C] //Proc of the 12th Int Conf on World Wide Web. Stroudsburg, PA: ACL, 2003: 519-528

[11] Baccianella S, Esuli A, Sebastiani F. Multi-facet Rating of Product Reviews [M]. Berlin: Springer, 2009: 461-472

[12] Qu L, Ifrim G, Weikum G. The bag-of-opinions method for review rating prediction from sparse text patterns [C] //Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2010: 913-921

[13] Wang H N, Yue L, Cheng X Z. Latent aspect rating analysis on review text data: A rating regression approach [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 783-792

[14] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis [C] //Proc of the 5th Joint Conf on Human Language Technology and the 10th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2005: 347-354

[15] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2): 1-135

[16] Elsner M, Austerweil J, Charniak E. A unified local and global model for discourse coherence [C] //Proc of Human Language Technologies: The 2007 Annual Conf of the 9th American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2007: 436-443

[17] Beineke P, Hastie Y, Manning C, et al. Exploring sentiment summarization [C] //Proc of the 9th National Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2004: 12-15

[18] Haghighi A, Vanderwende L. Exploring content models for multi-document summarization [C] //Proc of Human Language Technologies: The 2009 Annual Conf of the 9th American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2009: 362-370

[19] Blei D M, Ng A Y, Jordan M. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022

[20] Maaten L, Welling M, Saul L K. Hidden-unit conditional random fields [C] //Proc of the 14th Int Conf on Artificial Intelligence and Statistics. Stroudsburg, PA: ACL. 2011: 479-488

[21] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554

[22] Deng L, Yu D. Deep convex net: A scalable architecture for speech pattern classification [C] //Proc of the 2011 Interspeech. Stroudsburg. PA: ACL. 2011: 2285-2288

[23] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C] //Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 1096-1103

[24] Richardson M, Domingos P. Markov logic networks [J]. Machine Learning, 2006, 62(1/2): 107-136

[25] Kohonen T. Self-organizing Maps [M]. Berlin: Springer, 2001: 105-176

[26] Shi Chunqi, Shi Zhiping, Liu Xi, et al. Image segmentation based on self-organizing dynamic neural network [J]. Journal of Computer Research and Development, 2009, 46(1): 23-30 (in Chinese)
(史春奇, 施智平, 刘曦, 等. 基于自组织动态神经网络的图像分割[J]. 计算机研究与发展, 2009, 46(1): 23-30)

[27] Poon H, Domingos P. Unsupervised semantic parsing [C] //Proc of the 2009 Conf on Empirical Methods in Natural Language Proceesing. Stroudsburg, PA: ACL, 2009: 1-10

[28] Cheng Xingxin. The convergence of EM algorithm [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 1987, 23(3): 1-8 (in Chinese)
(程兴新. EM算法收敛性[J]. 北京大学学报: 自然科学版, 1987, 23(3): 1-8

[29] Daume III H, Marcu D. Domain adaptation for statistical classifiers [J]. Journal of Artificial Intelligence Reasearch, 2006, 26(1): 101-126

[30] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification [C] //Proc of the 45th Annual Meeting of the Assoc Computational Linguistics. Stroudsburg, PA: ACL, 2007: 432-439

Wu Lei, born in 1985. PhD candidate in the Institute of Automation, Chinese Academy of Sciences. Her research interests include probabilistic graphical model, machine learning.

Zhang Wensheng, born in 1966. Professor and PhD supervisor in the Institute of Automation, Chinese Academy o0f Sciences. His research interests include pattern recognition and machine learning, big data mining, probabilistic graphical model, deep learning, 3D numerical simulation and video image processing.

Wang Jue, born in 1948. Professor and PhD supervisor in the Institute of Automation, Chinese Academy of Sciences. His research interests include artificial intelligence, machine learning and data mining.

2015 年起《计算机研究与发展》双月将固定领域专题

致广大读者和作者:

本刊从 2015 年起将双数期约 1/2 版面固定为某个领域,每年将策划该领域的一个热点主题进行集中报道.具体的征文通知将在专题发表前 6 个月发布,请关注期刊网站!

此外,本刊依然欢迎自由来稿.谢谢!

具体领域分布及执行领域编委如下:

刊期	领域	领域编委	投稿方式	截稿日期
2 期	软件技术(含数据库)	孟小峰 xfmeng@ruc.edu.cn 中国人民大学	期刊网站投稿, 备注填写“年+专题名称”	上一年 10 月 1 日左右 (以具体征文通知为准)
4 期	网络技术	林闯 chlin@tsinghua.edu.cn 清华大学	期刊网站投稿, 备注填写“年+专题名称”	上一年 12 月 1 日左右 (以具体征文通知为准)
6 期	体系结构	刘志勇 zyliu@ict.ac.cn 中国科学院计算技术研究所	期刊网站投稿, 备注填写“年+专题名称”	当年 2 月 1 日左右 (以具体征文通知为准)
8 期	人工智能	周志华 zhouzh@nju.edu.cn 南京大学	期刊网站投稿, 备注填写“年+专题名称”	当年 4 月 1 日左右 (以具体征文通知为准)
10 期	信息安全	曹珍富 zfcao@sjtu.edu.cn 上海交通大学	期刊网站投稿, 备注填写“年+专题名称”	当年 6 月 1 日左右 (以具体征文通知为准)
12 期	应用技术	郑庆华 qhzheng@mail.xjtu.edu.cn 西安交通大学	期刊网站投稿, 备注填写“年+专题名称”	当年 8 月 1 日左右 (以具体征文通知为准)