·人工智能及识别技术· 文章编号：1000-3428(2016)08-0153-07 文献标志码：A 中图分类号：TP181

小样本贝叶斯网络参数学习方法

李子达，廖士中

（天津大学 计算机科学与技术学院，天津 300350）

摘 要： 当训练数据充分时，极大似然估计方法是贝叶斯网络参数学习典型且有效的方法。但当训练数据量少且领域知识缺乏时，极大似然估计往往无法给出一致无偏的参数估计。为此，提出一种新的贝叶斯网络参数学习方法 TL-WMLE。将极大似然估计方法与迁移学习理论、样本不均衡方法相结合，解决数据量过少、领域知识缺乏时的贝叶斯网络参数学习问题。使用 SMOTE-N 方法构建辅助分类器，并依据协变量偏移理论，利用辅助分类器的分类结果来计算源域数据权值。采用赋权的源域数据和目标域数据构造目标域的似然函数，应用该似然函数对目标域的参数进行极大似然估计。实验结果表明，在小样本情况下，该方法的分类精度优于极大似然估计方法。

关键词： 贝叶斯网络；参数学习；小样本；迁移学习；目标域

中文引用格式： 李子达，廖士中. 小样本贝叶斯网络参数学习方法[J]. 计算机工程，2016，42(8)：153-159，165.

英文引用格式： Li Zida，Liao Shizhong. Bayesian Network Parameter Learning Method on Small Samples[J]. Computer Engineering，2016，42(8)：153-159，165.

Bayesian Network Parameter Learning Method on Small Samples

LI Zida，LIAO Shizhong

（School of Computer Science and Technology，Tianjin University，Tianjin 300350，China）

【Abstract】 Maximum likelihood estimation is a classical and effective method for Bayesian network parameter learning on large samples，but it is not consistent when learning on small sample with little expertise. To address the issue，a novel method called TL-WMLE is proposed for Bayesian network parameter learning，which combines maximum likelihood，transfer learning and imbalance sample methods. The novel method uses an auxiliary classifier constructed by the SMOTE-N method and covariate migration theory，and computes the weights of source samples according to the predicted probability of the source domain by the auxiliary classifier. Then the proposed method mixes the reweighted source train sample and the target train sample to build a likelihood function on the target domain，and uses the new likelihood function to learn the parameters of the target domain via maximum likelihood estimation. Experimental results demonstrate that the classification accuracy of the proposed method outperforms that of the likelihood method on small samples.

【Key words】 Bayesian Network(BN)；parameter learning；small sample；transfer learning；target domain

DOI： 10.3969/j.issn.1000-3428.2016.08.028

1 概述

贝叶斯网络(Bayesian Network，BN)是图论与概率论的结合，为多元概率分布提供了一种紧凑的表示。贝叶斯网络由一个有向无环图(Directed Acyclic Graph，DAG)和该图中每个节点对应的条件概率表(Conditional Probability Tables，CPTs)组成。有向无环图中每个节点表示一个随机变量，边表示随机变量之间的依赖关系，条件概率表量化地表示依赖的程度。贝叶斯网络的学习包括结构学习和参数学习两部分，结构学习的目的是找到最符合数据特点的有向无环图。参数学习是指在结构已经具备的情况下，学习各个节点对应的条件概率表。

贝叶斯网络的性能很大程度上取决于参数学习的效果，即在给定结构情况下，每个节点条件概率表的学习对贝叶斯网络性能往往起到决定性作用[1-4]。通常情况下，参数学习最简单有效的方法是极大似然估计(Maximum Likelihood Estimation，MLE)。极

基金项目： 国家自然科学基金资助项目“机器学习核方法模型选择与组合的核矩阵近似分析方法”(61170019)。

作者简介： 李子达(1990－)，男，硕士研究生，主研方向为机器学习、数据挖掘；廖士中，教授、博士。

收稿日期： 2015-08-10 **修回日期：** 2015-09-26 **E-mail：** szliao@tju.edu.cn

大似然估计在数据充足的情况下可以得到对原始分布的无偏估计。然而,当数据量稀少并且领域知识缺乏时,极大似然估计的结果存在较大偏差。

以往,在数据量稀少的情况下,有 2 类基本的参数训练方法。第一类方法的前提是有充足的领域知识,通过将领域知识作为先验,使用贝叶斯估计来进行参数学习。但是,充足的领域知识往往很难被获取。第二类方法引入人工条件来约束贝叶斯网络的参数学习过程[2,5],这些人工条件可以是相应领域专家的意见,引入约束将参数学习变为有约束的优化问题。但是专家只能给出定性的意见,这样的意见融合进学习过程将不可避免地带来偏差[6]。另一些工作如文献[7]虽然引入了定量的约束条件,但是当贝叶斯网络参数数量较大时,人工加入的约束条件变得极为不充分[5],同时,过多的约束条件使得相应优化过程复杂难解。

本文利用转导迁移学习[8-9]中协变量偏移理论[10]改进贝叶斯网络参数学习的极大似然估计方法,使用赋权的源域数据样本和目标域数据样本构建目标域上的似然函数,利用样本不平衡分类方法的分类结果计算源域样本权值,提出一种在贝叶斯网络结构已知、训练数据稀少的情况下的贝叶斯网络参数学习方法。

2 贝叶斯网络及其参数学习方法

贝叶斯网络为多元随机变量 $x=\{X_1,X_2,\cdots,X_n\}$ 的概率分布提供了一种图表示,记做$\langle G,\theta\rangle$。其中,图 $G=\langle N,E\rangle$为有向无环图,图中节点 $N_i\in N,i=1,2,\cdots,n$,对应随机变量 X_i;图中边 $e_j\in E,j=1,2,\cdots,e_G$ 表示随机变量间的依赖关系,若 X_i 依赖于 X_j,则记 X_j 为 X_i 的父节点,X_i 的父节点集合记为 π_i;$\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 为一组条件概率分布表,其中 θ_i,$i=1,2,\cdots,n$,为随机变量 X_i 的条件概率分布表,X_i 有 R_i 种取值,其父节点集合取值格局记作 $\pi_i^k,k=1,2,\cdots,k_{\pi_i}$,则 $\theta_i\in\mathbf{R}^{R_i\times k_{\pi_i}}$中每一项可表示为:

$$P(X_i^r\mid\pi_i^k;\theta_i)=\theta_{irk}$$
$$r=1,2,\cdots,R_i$$
$$k=1,2,\cdots,k_{\pi_i} \tag{1}$$

贝叶斯网络具有模块化的性质,即每个随机变量及其父节点记为一个模块,每个随机变量的条件概率表和其他随机变量相互独立。同一随机变量不同父节点取值格局下,该随机变量取值的条件概率也相互独立,即:

$$P(x)=\prod_i^n P(X_i\mid\pi_i;\theta_i) \tag{2}$$

参数学习时按照模块分别学习相应参数。

极大似然估计是贝叶斯网络参数估计最常用的方法。用 m_{irk}表示随机变量 X_i 在数据中取第 r 种取值,且其父节点取值格局为 π_i^k 的样本个数。$D=\{D_1,D_2,\cdots,D_m\}$ 表示数据样本集。对每个模块,极大似然估计的对数似然函数表示如下:

$$\begin{aligned}l(\theta\mid D)&=\sum_{l=1}^m\ln P(D_l\mid\theta)\\&=\frac{1}{m}\sum_{l=1}^m\sum_{i=1}^n\sum_{r=1}^{R_i}\sum_{k=1}^{k_{\pi_i}}I(X_i^r,\pi_i^k\mid D_l)\\&\quad\cdot\ln P(X_i^r\mid\pi_i^k,\theta_i)\end{aligned}$$

其中,$I(X_i^r,\pi_i^k\mid D_l)$ 表示取值(X_i^r,π_i^k)出现在样本 D_l 中的指示函数;$P(X_i^r\mid\pi_i^k;\theta_i)$定义见式(1)。用 $\hat{\theta}_{irk}$表示 θ_{irk}的极大似然估计,其取值为:

$$\hat{\theta}_{irk}=\frac{m_{irk}}{\sum_{r=1}^{R_i}m_{irk}}$$

3 转导迁移学习与协变量偏移

转导学习的目的是依据采自分布 $P_S(x,y)$的样本,计算依目标域分布 $P_t(x,y)$最小化损失的最优参数 θ_t^*,即:

$$\theta_t^*=\arg\min_{\theta\in\Theta}\int_{x,y}P_t(x,y)l(x,y,\theta)\mathrm{d}(x,y)$$

转导迁移学习发生在源域数据与目标域数据分布不同、目标域数据较少、源域数据充分且目标域同源域存在联系的情况下。对于源域数据样本,$D=\{D_1,D_2,\cdots,D_{m_S}\}$,可利用源域数据构造目标域上的损失函数进行参数学习,即:

$$\begin{aligned}\theta_t^*&=\operatorname{argmin}_{\theta\in\Theta}\int_{x,y}\frac{P_t(x,y)}{P_s(x,y)}P_s(x,y)l(x,y,\theta)\mathrm{d}(x,y)\\&\approx\operatorname{argmin}_{\theta\in\Theta}\int_{x,y}\frac{P_t(x,y)}{P_s(x,y)}\tilde{P}_s(x,y)l(x,y,\theta)\mathrm{d}(x,y)\\&=\operatorname{argmin}_{\theta\in\Theta}\sum_{i=1}^{m_s}\frac{P_t(x_i^s,y_i^s)}{P_s(x_i^s,y_i^s)}l(x_i^s,y_i^s,\theta)\end{aligned}$$

其中,$\tilde{P}(x,y)$表示分布 $P(x,y)$的经验分布。对源域数据样本(x_i^s,y_i^s)的损失函数赋权值$\frac{P_t(x_i^s,y_i^s)}{P_s(x_i^s,y_i^s)}$,构造目标域上的损失函数,学习出目标域上的有效参数,改善目标域上的参数学习效果,这种方法叫作样本赋权。

文献[10]认为,源域分布与目标域分布中 y 相对于 x 的条件概率相同,即 $P_t(y\mid x)=P_s(y\mid x)$,但联合分布不同,即:

$$\begin{aligned}P_s(x,y)&=P(y\mid x)P_s(x)\\&\neq P_t(x,y)=P(y\mid x)P_t(x)\end{aligned}$$

这时,有$\frac{P_t(x,y)}{P_s(x,y)}=\frac{P_t(x)}{P_s(x)}$,源域数据和目标域数

据间的这种关系叫作协变量偏移(Covariate Shift)。

在协变量偏移的情况下,由源域训练出来的模型参数 θ_s 并不能直接用在目标域上,由于 x 依据源域分布,在模型集 $\{P(Y|X,\theta)\}_{\theta\in\Theta}$ 上可能不存在 θ 使得对所有 x 有 $P(Y|X=x,\theta)=P(Y|X=x)$,这种情况叫作模型特化错误(Miss-Specified)。

本文假设用于贝叶斯网络训练的目标域数据与源域数据间存在协变量偏移。

4　带权的极大似然估计参数学习方法

带权极大似然估计(Transfer Learning based Weighted Maximum Likelihood Estimation,TL-WMLE)分为 2 个阶段,第 1 阶段构造辅助分类器,区分目标域数据和源域数据,并利用分类概率计算源域数据的样本权值;第 2 阶段利用第 1 阶段计算出的权值、源域数据、目标域数据构建带权的似然函数,并计算带权极大似然估计。本节首先介绍带权极大似然函数的构建和带权极大似然估计的计算,然后介绍权值的计算。

4.1　基本假设

根据式(1),贝叶斯网络中每个模块的参数估计都相当于一个独立的条件概率估计问题,对贝叶斯网络中的每个模块 $\{X_i,\pi_i\}$,其在源域中和目标域中的分布分别记为:

$$P_S(X_i,\pi_i)=P_S(X_i|\pi_i)P_S(\pi_i)$$

$$P_T(X_i,\pi_i)=P_T(X_i|\pi_i)P_T(\pi_i)$$

利用协变量偏移理论,假设如下:

假设 1　在源域中与目标域中,随机变量 X_i 相对于其父节点集 π_i 的条件概率分布相同,即 $P_S(X_i|\pi_i)=P_T(X_i|\pi_i)=P_i(X_i|\pi_i)$,即贝叶斯网络中每个模块都存在协变量偏移。

假设 2　源域分布 $P_S(x)$ 与目标域分布 $P_T(x)$ 支持相同或在很大程度上重叠。

满足上述 2 条假设的源域分布和目标域分布存在协变量偏移。直接使用样本赋权方法需要对贝叶斯网络中每个模块单独计算权重,需要的权重个数为样本数乘以模块数,计算量大。

4.2　带权的极大似然估计

定义如下损失函数:

$$Loss(J)=-\int P_J(x)\sum_i\sum_k\ln P(X_i|\pi_i)\mathrm{d}x \tag{3}$$

其中,$J\in\{S,T\}$,分别表示源域和目标域上的损失。对数据样本 $D=\{D_1,D_2,\cdots,D_{m_J}\}$,记似然函数如下:

$$\begin{aligned}l_J(\theta|D)&=\frac{1}{m_J}\sum_{l=1}^{m_J}\ln P(D_l|\theta)\\&=\frac{1}{m_J}\sum_{l=1}^{m_J}\sum_{i=1}^{n}\sum_{r=1}^{R_i}\sum_{k=1}^{k_{\pi_i}}f(X_i^r,\pi_i^k|D_l,\theta_i)\end{aligned} \tag{4}$$

其中,$f(X_i^r,\pi_i^k|D_l,\theta_i)=I(X_i^r,\pi_i^k|D_l)\ln P(X_i^r|\pi_i^k;\theta_i)$ 表示属性 X_i 在数据 D_l 中的似然函数,$I(X_i^r,\pi_i^k)$ 表示随机变量 X_i 及其父节点的取值格局 (X_i^r,π_i^k) 出现在样本 D_l 中的指示函数,若出现,则 $I(X_i^r,\pi_i^k)=1$,否则 $I(X_i^r,\pi_i^k)=0$。存在:

$$\lim_{m\to\infty}-l_J(\theta|D)=Loss(J) \tag{5}$$

记 $\hat{\theta}_J=\mathrm{argmax}\,l_J(\theta|D)$ 为域 J 上的极大似然估计,$\theta_J^*=\mathrm{argmin}\,Loss(J)$ 为域 J 上损失函数最小化参数,由上式可以看出,当 $m\to\infty$ 时,有 $\hat{\theta}_J$ 依概率趋近于 θ_J^*。

对于目标域,结合式(3)~式(5),以及假设 1,可知源域上的极大似然估计与目标域上的极大似然估计之间存在联系。对于目标域上的似然函数 $l_T(\theta|D)$,有:

$$\begin{aligned}\lim_{m_T\to\infty}-l_T(\theta|D)&=Loss(T)\\&=-\int P_T(x)\sum_i\sum_k\ln P(X_i|\pi_i)\mathrm{d}x\\&=-\int P_S(x)\frac{P_T(x)}{P_S(x)}\sum_i\sum_k\ln P(X_i|\pi_i)\mathrm{d}x\\&\approx\frac{1}{m_S}\sum_{l=1}^{m_S}\sum_{i=1}^{n}\sum_{r=1}^{R_i}\sum_{k=1}^{k_{\pi_i}}\frac{P_T(D_l)}{P_S(D_l)}\\&\quad\cdot f(X_i^r,\pi_i^k|D_l,\theta_i)\end{aligned} \tag{6}$$

由式(6)可以看出,当数据样本数趋于无穷时,源域数据上的带权极大似然估计,依概率趋于目标域上的极大似然估计。所以,可以将目标域与源域数据混合,为源域数据样本赋权 $\frac{P_T(D_l)}{P_S(D_l)}$,目标域数据样本赋权 1,构建如式(7)所示的带权的极大似然估计:

$$\begin{aligned}\hat{\theta}^w&=\arg\max_\theta\frac{1}{m_S}\sum_{l=1}^{m_S}\sum_{i=1}^{n}\sum_{r=1}^{R_i}\sum_{k=1}^{k_{\pi_i}}\frac{P_T(D_l)}{P_S(D_l)}f(X_i^r,\pi_i^k|D_l,\theta_i)\\&\quad+\frac{1}{m_T}\sum_{l=1}^{m_T}\sum_{i=1}^{n}\sum_{r=1}^{R_i}\sum_{k=1}^{k_{\pi_i}}f(X_i^r,\pi_i^k|D_l,\theta_i)\end{aligned} \tag{7}$$

统一表示目标域和源域的权重为 $\phi(D_l)$:当 D_l 为源域训练样本,$\phi(D_l)=\frac{P_T(D_l)}{P_S(D_l)}$;当 D_l 为目标域训练样本,$\phi(D_l)=1$。对源域和目标域训练样本,式(7)表示为:

$$\hat{\theta}_{irk}^w=\frac{\sum_{l=1}^{m_s+m_T}\phi(D_l)I(X_i^r,\pi_i^k|D_l)}{\sum_{l=1}^{m_s+m_T}\sum_{u=1}^{R_i}\phi(D_l)I(X_i^u,\pi_i^k|D_l)} \tag{8}$$

式(8)依照 K-L 距离的定义即可求得[11]。从式(7)、式(8)可以看出,该方法不需要领域专家知识,不引入复杂难解的优化过程,是一种简单而有效的方法。目标域权值恒定为 1。

4.3 权值计算

通常情况下，$\phi(D_l)=\frac{P_T(D_l)}{P_S(D_l)}$可以通过分别估计源域数据分布和目标域数据分布得到。但是，分布估计的方法存在如下问题：首先，分布估计方法普遍偏差较大且复杂度高[12]；其次，当目标域样本数过少时，由于数量可能远小于分布估计的有效数据量，对目标域进行的分布估计会远远偏离目标域的实际分布，在小样本情况下，分布估计的方法不可行。文献[12]提出了一种方法，将分布估计转化为分类来解决分布估计存在的问题。该方法引入辅助分类器对某一样本属于源域或目标域进行分类，并得到该样本属于源域或目标域的分类概率，通过所得分类概率来计算分布比值 $\phi(D_l)$。针对小样本贝叶斯网络参数学习问题，TL-WMLE 将该方法和样本不均衡分类方法相结合，不对源域和目标域样本进行分布估计，通过样本不均衡分类方法构建辅助分类器，辅助分类器的分类概率用于计算分布比值 $\phi(D_l)$，避免了小样本上进行分布估计所带来的偏差。

依据文献[12]，对每个源域和目标域中的样本，定义选择变量 σ，$\sigma=1$ 表示样本来自源域，$\sigma=0$ 表示样本来自目标域。引入参数 λ，φ，表示源域和目标域分布为 $P_T(x)=P(x|\varphi)$，$P_S(x)=P(x|\lambda)$，则有：

$$\frac{p_T(x)}{p_S(x)}=\frac{P(\sigma=1|\varphi,\lambda)P(\sigma=0|\varphi,\lambda)p(x|\varphi)}{P(\sigma=0|\varphi,\lambda)P(\sigma=1|\varphi,\lambda)p(x|\lambda)}$$
$$=\frac{P(\sigma=1|\varphi,\lambda)}{P(\sigma=0|\varphi,\lambda)}\left(\frac{P(\sigma=0|\varphi,\lambda)p(x|\varphi)}{P(\sigma=1|\varphi,\lambda)p(x|\lambda)}\right)$$

上式中等号右边括号内，有：

$$\frac{P(\sigma=0|\varphi,\lambda)P(x|\varphi)}{P(\sigma=1|\varphi,\lambda)P(x|\lambda)}=1+\frac{P(\sigma=0|\varphi,\lambda)P(x|\varphi)}{P(\sigma=1|\varphi,\lambda)P(x|\lambda)}-1$$
$$=\left(\frac{P(\sigma=1|\varphi,\lambda)P(x|\lambda)}{P(\sigma=1|\varphi,\lambda)P(x|\lambda)+P(\sigma=0|\varphi,\lambda)P(x|\varphi)}\right)^{-1}-1$$
$$=\frac{1}{P(\sigma=1|x,\varphi,\lambda)}-1$$

最后一步推导使用了贝叶斯定理，带入得：

$$\frac{P_T(x)}{P_S(x)}=\frac{P(\sigma=1|\varphi,\lambda)}{P(\sigma=0|\varphi,\lambda)}\left(\frac{1}{P(\sigma=1|x,\varphi,\lambda)}-1\right) \tag{9}$$

其中，$p(\sigma=1|x,\varphi,\lambda)$ 表示样本 x 属于源域数据的概率，该概率通过辅助分类器得到；$\frac{P(\sigma=1|\varphi,\lambda)}{P(\sigma=0|\varphi,\lambda)}$表示源域训练样本和目标域训练样本数目的比值。

由式(9)可以构造一个辅助分类器，对混合的目标域训练样本和源域训练样本进行分类，利用分类概率计算样本 x 的权值 $\phi(x)$。本文使用朴素贝叶斯分类器作为辅助分类器。由于目标域样本数据较少，目标域数据数目与源域数据数目不均衡，直接构造辅助分类器可能存在偏差，本文方法中引入样本不均衡分类方法构造辅助分类器。

参照文献[13]，由于朴素贝叶斯分类器预测出的概率值会比较靠近 0 或 1，因此对上述比值进行规则化，引入规则化系数 α，规则化后的比值为：

$$\frac{P_T(x)}{P_S(x)}\propto\frac{1}{1+e^{-\alpha\ln\frac{1-P(\sigma=1|x,\lambda,\varphi)}{P(\sigma=1|x,\lambda,\varphi)}}} \tag{10}$$

使用式(10)中右半部分作为源域数据权值。

4.4 辅助分类器构造

辅助分类器的主要作用是对混合的源域数据和目标域数据进行区分(分类)，并依据式(10)，使用分类概率 $p(\sigma=1|x,\varphi,\lambda)$ 计算样本权值。样本不平衡是指在样本中一类样本的数量远多于另一类样本的数量。一般情况下，分类器适用于样本数量均衡的情况，不适用于样本数量不均衡的情况[14]。源域数据远多于目标域数据，导致辅助分类器面临样本不均衡问题，需采用样本不均衡分类方法构建辅助分类器。

辅助分类器按如下过程构造，首先使用对于标称数据的综合少数过采样技术(Synthetic Minority Over-sampling Technique，SMOTE-N)[14]来增加目标域训练样本。SMOTE-N 根据数据特征空间相似性为目标域数据增加人工样本，其过程如下：

输入 目标域数据集

输出 人工数据样本集

对目标域数据中每个样本：

(1)根据值差异度量(Value Difference Metric，VDM)[14]作为样本距离度量，计算目标域数据样本中的 k 近邻。

(2)新样本中，对每一个属性，取该属性在 k 个近邻中出现最多的取值。

之后，将得到的人工数据样本集与目标域训练样本混合，得到新的目标域训练样本 D_{new}，为 D_{new} 增加一维，作为新的类别标志，取值为 0，为源域训练样本增加一维，取值为 1，混合来构建朴素贝叶斯分类器。则该分类器分类为 1 的概率即为所求 $p(\sigma=1|x,\varphi,\lambda)$。

5 实验结果与分析

实验的目的在于验证 TL-WMLE 算法在数据较

少情况下进行贝叶斯网络参数学习的有效性。实验总共分为两部分,第 1 部分为模拟实验,模拟实验分为 2 组,模拟实验中通过人工方法产生目标域和源域数据,通过对比目标域测试数据上的似然或分类准确率来比较 TL-WMLE 与目标域上极大似然估计(MLE)。第 2 部分实验为实际应用实验,在文本情感分类数据集上通过 TL-WMLE 和极大似然估计分别构建朴素贝叶斯分类器,利用目标域测试数据集上的分类准确率进行对比。实验环境为 8 GB 内存,2.8 GHz Intel i5 中央处理器。

5.1　模拟实验设置与实验结果

模拟实验分为 2 组,分别从平均似然估计和分类准确率的角度对比 TL-WMLE 和极大似然估计(MLE)。模拟实验中源域训练数据和目标域训练数据为人工构造,构造方法参照文献[15],方法如下:对采样自同一分布 $P(x)$ 的全部数据,用 $\sigma=1$ 表示某数据被选为目标域数据。对于某个样本 x 中的某个属性 X_i,定义 $P(\sigma=1|X_i\in r,x)=0.2$,表示当 X_i 取 r 中值时该样本被选为目标域的概率为 0.2;$P(\sigma=1|X_i\notin r,x)=0.8$ 表示该属性取值不在 r 中时该样本被选为目标域数据概率为 0.8。选中的数据作为目标域数据,未被选中的数据作源域数据,则源域数据和目标域数据服从不同分布,源域数据服从分布 $P(x)P(\sigma=0|x)$,目标域数据服从分布 $P(x)P(\sigma=1|x)$。将目标域数据随机划分成目标域训练数据和目标域测试数据。

5.1.1　贝叶斯网络模拟实验的人工构建

第 1 组模拟实验中,构造简单贝叶斯网络,从该网络中采集数据,将所采集的数据划分为源域数据和目标域数据,使用 TL-WMLE 与极大似然估计分别训练贝叶斯网络参数,对比所得参数在目标域测试数据上的平均似然估计。

构造拥有 5 个节点的贝叶斯网络,结构如图 1 所示。每个节点取值属于{0,1},并设置节点的条件概率表,如表 1 所示。

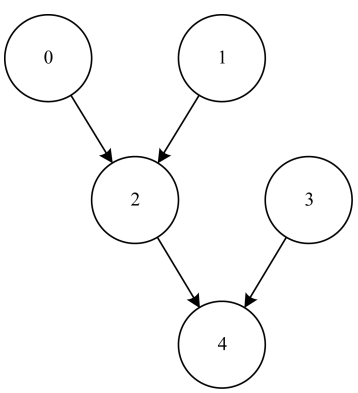

图 1　第 1 组模拟实验贝叶斯网络结构

表 1　第 1 组模拟实验贝叶斯网络条件概率

节点	节点取值	父节点格局	条件概率
0	0	–	0.55
1	0	–	0.68
	0	0,0	0.50
2	0	0,1	0.60
	0	1,0	0.70
3	0	1,1	0.70
	0	–	0.48
	0	0,0	0.43
4	0	0,1	0.78
	0	1,0	0.50
	0	1,1	0.40

从该网络中用向前采样方法采集 1 000 个 ~ 3 000 个数据样本。用数据划分方法进行源域数据和目标域数据划分,可得到大约 5% 的数据作为目标域数据,将 2% 作为目标域训练数据,3% 作为目标域测试数据,其余 95% 数据作为源域数据。规则化参数 $\alpha=1$。

通过计算目标域测试数据集上样本的平均似然估计大小,来比较目标域训练数据上的极大似然估计与 TL-WMLE,每种数据量情况下重复 100 次,测试数据的平均似然估计记录如表 2 所示。

表 2　目标域测试数据样本的平均似然估计

数据总数	TL-WMLE	MLE
1 000	-0.719	-0.720
1 500	-0.719	-0.738
2 000	-0.709	-0.712
2 500	-0.671	-0.707
3 000	-0.704	-0.707

从表 2 可以看出,随着训练数据量的提高,目标域训练数据上的极大似然估计所得模型在目标域测试数据上的平均似然估计有所提高,这符合极大似然估计的性质[10]。TL-WMLE 训练得到的贝叶斯网络模型在目标域测试数据上的平均似然估计均高于由极大似然估计在目标域训练数据上所得的贝叶斯网络模型,说明 TL-WMLE 相比于极大似然估计所得模型更接近目标域实际分布。2 种方法所得似然差异逐渐减小是因为训练数据量逐渐增大,使得目标域训练数据逼近有效数据量[10]。

5.1.2　标准数据集上的朴素贝叶斯分类实验

第 2 组实验在标准数据集上进行。所用数据集均来自 UCI 机器学习数据集。使用的数据集如表 3 所示。使用 TL-WMLE 与极大似然估计分别构建朴素贝叶斯分类器,通过目标域测试数据上的分类准确率,比较 TL-WMLE 与极大似然估计。

表 3 第 2 组模拟实验所用标准数据集

数据集名称	数据维度	样本个数
Car Evaluation	6	1 728
Chess	36	3 196
Mushroom	22	8 124
Nursery	8	12 960
Connect-4	42	67 557

对每个数据集,分别随机取其中的 20% ,40% ,60% ,80% ,100% 作为实验样本。在实验样本中,使用 5.1 节中的划分方法,划分出源域训练数据与目标域训练数据、目标域测试数据。实验样本中,大约有 20% 的数据被划分为目标域数据,并将其中的 5% 作为目标域训练数据,另外 15% 作为目标域测试数据。参数 $\alpha = 1$。每种数据量情况下实验重复 5 次。实验结果如图 2 ~ 图 6 所示。

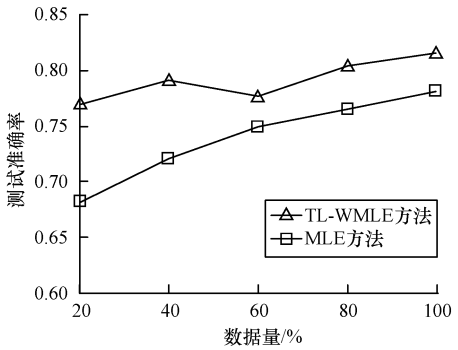

图 2 Car Evaluation 数据集上实验结果

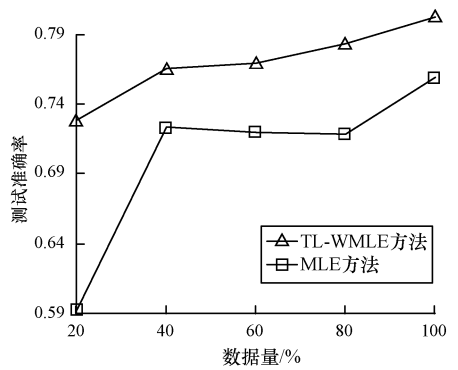

图 3 Chess 数据集上实验结果

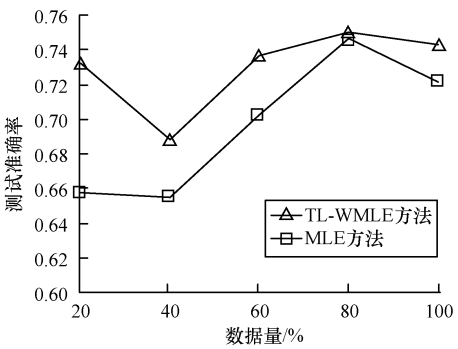

图 4 Mushroom 数据集上验结果

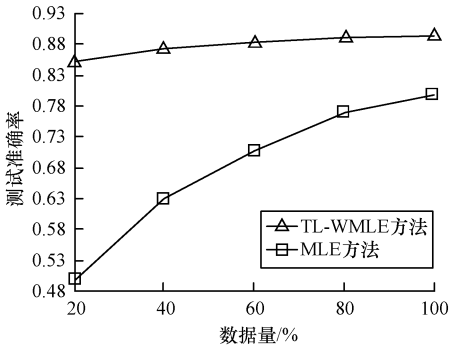

图 5 Nursery 数据集上实验结果

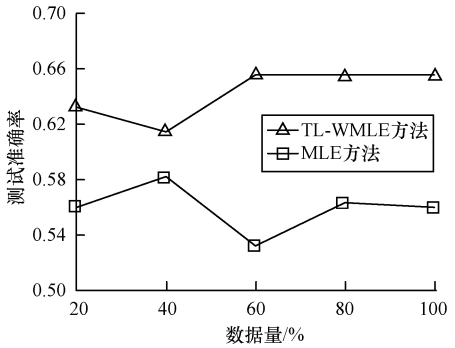

图 6 Connect-4 数据集上实验结果

在第 2 组模拟实验中,除 Connect-4 数据集外,其余 4 组极大似然估计的测试准确率均随数据量增长而增长,TL-WMLE 在 5 组数据上均明显好于极大似然估计,特别是对 Mushroom 数据集当目标域数据量小时,TL-WMLE 比极大似然估计性能提升超过 20% 。

5.2 实际应用实验

在文本情感分类数据集上用 TL-WMLE 与极大似然估计分别构建朴素贝叶斯分类器,通过对比在目标域测试数据上的分类准确率,验证 TL-WMLE 算法的有效性。

数据集来自约翰霍普金斯大学的文本情感分类数据集[13]。数据集是购物网站实际商品的评价记录,共包含 4 类商品,分别是 book, electronics, kitchen,dvd。每类数据含有记录 2 000 条。数据集每条记录表示一个文本,每条记录含有该文本中出现的单词、双词以及词频。每个文本属于“积极”或者“消极”其中一类。

实验使用 book 数据集作为源域,另外 3 个数据集分别作为目标域。在每个目标域数据集上,实验分为 3 组,分别随机取 50 个、100 个、200 个目标域数据作为目标域训练数据,其余数据作为目标域测试数据。每组中,源域数据分别取 200 个、400 个、600 个、800 个、1 000 个、1 600 个、2 000 个作为源域训练数据。规则化参数 α 通过格搜索产生,搜索范围为[-5,5],步长为 0.5。

对比方法包括:方法 A,仅使用目标域训练数据进行极大似然估计;方法 B,仅使用源域训练数据进行极大似然估计;方法 C,使用目标域训练数据和源域训练数据混合进行极大似然估计;方法 D,仅使用源域数据进行 TL-WMLE,即去掉式(8)中目标域数据的求和项;方法E,使用源域数据和目标域数据进行 TL-WMLE,即式(8)。方法 A ~ 方法 C 均属于极大似然估计(MLE),但是使用的训练数据不同。由于 3 组实验结果趋势近似和篇幅限制,现展示目标域训练数据数目为 100 时的学习结果。在目标域测试数据上的测试准确率如表 4 所示,其中,加粗的项表示最优的项;*nn* 表示源域训练数据集大小。

表 4　文本情感分类数据上的目标域测试准确率　%

迁移过程	训练方法	*nn* = 200	*nn* = 400	*nn* = 600	*nn* = 800	*nn* = 1 000	*nn* = 1 600	*nn* = 2 000
book 迁移到 dvd	A	63.85 ±1.40	68.07 ±0.70	67.20 ±0.1	70.58 ±0.14	**70.90 ±0.80**	70.82 ±0.14	71.93 ±0.14
	B	66.59 ±3.00	62.79 ±0.40	65.64 ±0.95	64.69 ±0.32	59.06 ±1.60	65.88 ±0.32	66.01 ±0.70
	C	68.97 ±0.48	**69.39 ±0.16**	70.13 ±0.20	71.77 ±0.17	70.40 ±0.75	72.54 ±0.71	71.53 ±0.11
	D	59.75 ±0.10	68.97 ±0.12	68.52 ±0.17	71.90 ±0.24	69.34 ±1.40	72.09 ±0.14	71.14 ±0.60
	TL-WMLE	**69.63 ±0.11**	67.57 ±0.15	**70.72 ±0.14**	**71.93 ±0.57**	70.69 ±0.55	**72.11 ±0.16**	**72.38 ±0.20**
book 迁移到 electronics	A	55.40 ±0.25	58.37 ±2.20	60.06 ±1.40	61.96 ±0.24	62.88 ±2.00	63.06 ±1.90	62.30 ±0.26
	B	64.57 ±0.15	63.01 ±7.90	61.93 ±5.20	68.30 ±1.10	70.33 ±0.29	67.78 ±1.40	69.07 ±0.13
	C	64.07 ±4.30	63.56 ±6.70	67.78 ±0.14	65.51 ±0.14	67.38 ±0.07	67.57 ±0.10	65.46 ±0.20
	D	0.5577 ±0.33	0.6240 ±0.31	0.6148 ±2.00	62.06 ±0.36	63.38 ±0.22	64.54 ±0.93	63.22 ±0.95
	TL-WMLE	**66.36 ±2.80**	**65.30 ±0.74**	**68.99 ±1.00**	**69.28 ±3.30**	**71.10 ±2.00**	**69.91 ±0.23**	**70.41 ±0.56**
book 迁移到 kitchen	A	64.54 ±0.14	64.57 ±0.23	63.12 ±0.01	65.47 ±0.26	64.49 ±0.20	65.65 ±0.11	65.47 ±0.67
	B	66.97 ±3.90	68.29 ±0.56	66.92 ±8.30	68.23 ±0.55	70.87 ±0.04	67.29 ±1.70	69.37 ±0.02
	C	**70.63 ±1.40**	68.63 ±0.81	68.00 ±1.00	70.66 ±0.01	67.89 ±0.17	68.26 ±0.0051	69.32 ±0.92
	D	65.71 ±8.50	66.73 ±0.24	63.81 ±0.22	67.05 ±0.14	66.55 ±0.67	68.00 ±0.14	68.08 ±3.30
	TL-WMLE	70.18 ±3.40	**71.16 ±0.25**	**70.37 ±1.50**	**73.72 ±0.36**	**71.85 ±0.37**	**70.50 ±0.33**	**74.43 ±0.32**

实验结果表明:(1)仅使用目标域训练数据进行极大似然估计(方法 A)的结果大部分较差,表明极大似然估计在数据量较少时不能得到一致的参数估计;(2)仅使用源域数据进行极大似然估计(方法 B)所得模型在目标域测试数据集上的测试性能不佳,甚至测试准确率有时不如仅用目标域训练数据训练得到的分类器的分类效果,表明由于源域和目标域数据分布不同,存在模型特化错误;(3)使用目标域数据和源域数据混合进行极大似然估计(方法 C)效果稍好于仅使用源域数据的 TL-WMLE(方法 D),两者均优于方法 A、方法 B;(4)使用 TL-WMLE 构建的朴素贝叶斯分类器的分类性能在绝大多数情况下为 5 种方法中最优,而且明显优于方法 A、方法 B,相比于将目标域数据同源域数据混合进行极大似然估计的方法 C,性能也有较大提升。

由实验结果及分析可知,TL-WMLE 为数据量较少、领域知识缺乏的情况下贝叶斯网络参数学习提供了一种新思路和有效解决途径。

6　结束语

本文结合协变量偏移理论和样本不均衡分类方法,提出一种基于样本赋权的极大似然估计方法。实验结果表明,该方法不需要引入领域专家的专门知识,优化简单,且易于实现,是一种可行且有效的小样本贝叶斯网络参数学习方法。

下一步将深度挖掘源域和目标域分布间的关系,考虑贝叶斯网络的敏感性、非目标节点对目标节点的影响程度,以改善源域数据权值的计算,并提升本文方法的学习效果。

参考文献

[1] Heckerman D, Dan G, David M C. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data[J]. Machine Learning, 1995, 20(3): 197-243.

[2] Grossman D, Domingos P. Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood[C]// Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM Press, 2004: 46-53.

[3] Shen Bin, Su Xiaoyuan, Greiner R, el al. Discriminative Parameter Learning of General Bayesian Network Classifiers[C]//Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence. Washington D. C., USA: IEEE Press, 2003: 296-305.

[4] 廖学清,吕　强,单冬冬. 数据缺失下学习贝叶斯网的 S-EM 算法[J]. 计算机工程,2009,35(8):214-216,219.

[5] Zhou Yun, Fenton N, Neil M, et al. Incorporating Expert Judgement into Bayesian Network Machine Learning[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: IJCAI Inc., 2013: 3249-3250.

(下转第 165 页)

6　结束语

本文提出一种面向军事文本信息的特征词向量描述方法,采用先优化分词结果再筛选特征词的思路构建特征词向量。通过命名实体识别和扩充词典库,优化军事文本信息的分词结果,利用领域特征词与常用词汇在测试领域和其他领域出现频次的差异,改进领域特征词筛选算法,剔除误选常用词汇。实验结果表明,本文方法所生成的特征词向量简洁、完备,能够识别出军事文本中的命名实体,领域特征词筛选效果明显。在此基础上,下一步工作可以针对特征词之间的语义问题,通过构建领域本体,找出特征词之间的同义、包含、从属等语义关系,降低向量维度。

参考文献

[1] 杨杰明,刘元宁,曲朝阳,等. 文本分类中基于综合度量的特征选择方法[J]. 吉林大学学报:理学版,2013,51(5):887-893.

[2] 吴海燕. 基于互信息与词语共现的领域术语自动抽取方法研究[J]. 重庆邮电大学学报:自然科学版,2013,25(5):690-693.

[3] 李江华,时　鹏,胡长军. 一种适用于复合术语的本体概念学习方法[J]. 计算机科学,2013,40(5):168-172.

[4] 傅　鹏,黄利强,付春雷 一种改进的面向文本的领域概念筛选算法[J]. 计算机科学,2012,39(6):253-256.

[5] Velardi P,Missikoff M,Basili R. Identification of Relevant Terms to Support the Construction of Domain Ontologies [C]//Proceedings of the Workshop on Human Language Technologies and Knowledge Management. New York,USA:ACM Press,2001:1-8.

[6] Sundheim B M. Named Entity Task Definition [C]// Proceedings of the 6th Message Understanding Conference. Berlin,Germany:Springer,1995:319-332.

[7] 张冰怡,魏　博,陈建成,等. 基于对偶编码的中文分词算法[J]. 南京理工大学学报,2014,38(4):526-530.

[8] 杨晓冬,邵根富. 基于本体的作战文书分词的关键技术研究[D]. 杭州:杭州电子科技大学,2013.

[9] 张广军,王建宁. 基于 XML 作战文书理解关键技术研究[D]. 南京:南京理工大学,2009.

[10] 杨　森. 军事文献中复杂字母词语的形式分析[J]. 社会纵横,2010(3):315-316.

[11] 张海泉. 武器家谱[J]. 当代军事文摘,2005(3):21-22.

[12] 赵　伟,夏庆锋. 一种基于有限状态自动机的多鱼协作顶球算法[J]. 兵工自动化,2012,31(11):59-62.

[13] 王惠仙,龙　华. 基于改进的正向最大匹配中文分词算法研究[J]. 贵州大学学报:自然科学版,2011,28(5):112-115.

[14] Navigli R,Velardi P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites[J]. Computational Linguistics,2004,30(2):151-179.

[15] 贾秀玲,文敦伟. 面向文本的本体学习中概念提取及关系提取的研究[D]. 长沙:中南大学,2007.

[16] 邱　哲,符滔滔,王学松. 开发自己的搜索引擎 Lucene + Heritrix[M]. 2 版. 北京:人民邮电出版社,2005.

[17] 张玉芳,杨　芬,熊忠阳,等. 基于上下文的领域本体概念和关系的提取[J]. 计算机应用研究,2010,27(1):74-76.

编辑　顾逸斐

(上接第 159 页)

[6] Druzdel M J, Van D G L C. Building Probabilistic Networks:"Where Do the Numbers Come From?"[J]. IEEE Transactions on Knowledge and Data Engineering, 2000,12(4):481-486.

[7] Zhou Yun, Fenton N, Neil M. Bayesian Network Approach to Multinomial Parameter Learning Using Data and Expert Judgments [J]. International Journal of Approximate Reasoning,2014,55(5):1252-1268.

[8] Pan Jialin, Yang Qiang. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1345-1359.

[9] 张建军,王士同,王　骏. 迁移学习数据分类中的 ESVM 算法[J]. 计算机工程,2012,38(8):173-176.

[10] Shimodaira H. Improving Predictive Inference Under Covariate Shift by Weighting the Log-likelihood Function[J]. Journal of Statistical Planning and Inference, 2000,90(2):227-244.

[11] 张连文,郭海鹏. 贝叶斯网引论[M]. 北京:科学出版社,2006.

[12] Bickel S, Brückner M, Scheffer T. Discriminative Learning for Differing Training and Test Distributions[C]//Proceedings of the 24th International Conference on Machine Learning. New York, USA: ACM Press,2007:81-88.

[13] Xia Rui,Hu Xuelei,Lu Jianfeng,et al. Instance Selection and Instance Weighting for Cross-domain Sentiment Classification via PU Learning[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: IJCAI Inc., 2013: 2176-2182.

[14] Chawla N V,Bowyer K W,Hall L O,et al. SMOTE: Synthetic Minority Over-sampling Technique [J]. Journal of Artificial Intelligence Research,2002,16(1): 321-357.

[15] Huang Jiayuan, Gretton A, Borgwardt K M, el al. Correcting Sample Selection Bias by Unlabeled Data[C]// Proceedings of NIPS'06. Vancouver, Canada: [s. n.], 2006:601-608.

编辑　刘　冰