**Semi-supervised Stochastic Multi-Domain Learning using Variational Inference**

会议：ACL2019

代码：<https://github.com/lrank/Code_VariationalInference-Multidomain>

**摘要**

NLP的监督模型依赖大量文本集，这些文本集与预期的测试设置非常相似。 不幸的是，匹配的文本通常没有足够的数量，而且，在文本的任何域内，数据通常是高度异构的。 在本文中，我们提出了一种方法，该方法使用潜在变量模型提取重要域信号作为多域学习系统的一部分，在该潜变量模型中，神经模型的各个部分基于推断的域进行随机门控。 我们比较了离散的和连续的潜在变量的使用情况，这些变量在域监督或域半监督设置下运行，其中仅对训练输入的子集才知道域。我们证明，与竞争基准域自适应方法（包括使用对抗学习的方法）相比，我们的模型可显着提高性能。

**1简介**

文本语料库通常是从几种不同的来源（例如新闻，文学，微博和网络爬网）进行整理的，这带来了从异类数据学习NLP系统的问题，以及此类模型如何良好地转移到测试设置中。从这些语料库中学习需要可以推广到不同领域的模型，这就是被称为转移学习或领域适应的问题（Blitzer等人，2007; Daum´e III，2007; Joshi等人，2012; Kim等人，2016 ）。 在大多数最新的框架中，模型都完全了解训练数据中实例的领域，并且该领域被视为离散的指标变量。 但是，实际上，数据通常很混乱，域标签并不总是可用，或者提供的文本样式和体裁信息有限。 例如，网络爬虫的语料库由各种形式的文本组成，例如新闻，市场营销，博客，小说和食谱，但是通常不会明确指定每个文档的类型。 此外，即使标有特定领域的语料库本身也可能是更特定领域的实例，例如，“新闻”文章将涵盖政治，体育，旅行，舆论等。准确地建模这些类型的数据需要了解每个实例的特定域以及每个测试实例的域，这对于以前未见的域中的测试数据尤其有问题。

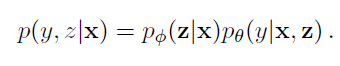
一种简单的领域学习策略是使用单个模型联合学习所有数据，其中模型不以领域为条件，并直接最大化p（y|x），其中x是文本输入，而y是输出（例如分类）标签。多域学习中报告的改进（DauméeIII，2007； Kim等人，2016）通常集中于学习每个实例的双胞胎表示（共享和私有表示）。 通过引入以域为条件的特定于域的通道来对专用表示进行建模，并通过域通用通道来学习共享的通道。 为了学习更强大的领域通用和领域特定渠道，可以采用领域条件方法或领域生成方法的形式来应用对抗监督（Liu等人，2016; Li等人，2018a）。

受到这些工作的启发，我们开发了一种设置区域的方法，其中未观察到或部分观察到域，就域而言，我们分别将其称为无监督和半监督。 通过将每个测试实例建模为多个领域的混合，这具有额外的优势，即可以从未见的领域中提取测试数据，从而提供鲁棒性。 在本文中，我们通过建模判别式学习问题，提出了使用潜在变量表征域的方法，其中z编码域，当该域未被观察到时必须将其边缘化。我们提出了一系列模型，这些模型在处理z的过程中越来越复杂，从离散的混合模型到连续的矢量值潜在变量（类似于主题模型； Blei等人（2003年）），使用Beta或Dirichlet分布建模。 我们展示了如何使用基于直接梯度的方法或变分推论（Kingma等，2014）针对各个模型类型有效地训练这些模型。 变分方法可以应用于域和/或标记半监督的设置，在该设置中，不能完全观察到训练数据的所有组成部分。

我们通过对多域产品评论数据和来自不同域的7种语言识别基准进行情感分析来评估我们的方法，这表明在域外评估中，我们的方法大大优于基准方法，包括对抗训练的域自适应（Li等，2018a）。 我们显示，包括其他域未标记的数据可以极大地提高性能，从而导致就我们所知，传输模型的性能通常优于域训练的模型，从而为数据集设置了新的技术水平。

**2随机域自适应**

在本节中，我们描述了我们提出的随机域自适应（SDA）方法，该方法使用潜在变量表示隐式“域”。 这被公式化为输出分类标签y和潜域z的联合模型，它们都以x为条件，



这两个分量分别是先验和分类器似然性，分别由进行参数化。 基于z的性质，我们建议先验的几种不同选择，即是否为：（i）离散值（“ DSDA”，请参见第2.2节）； 或（ii）连续向量，在这种情况下，我们将对模型进行不同的分布实验（“ CSDA”，请参见第2.3节）。

**2.1随机通道门控**

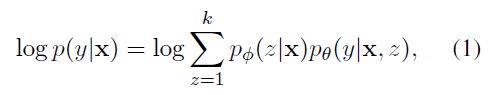
对于我们所有的模型，似然被表述为多通道神经模型，其中z被用作门来选择应该使用哪些通道来代表输入。 该模型包含k个通道，每个通道计算一个独立的隐藏表示，



使用卷积神经网络，我们的方法是通用的，可以很容易地与CNN之外的其他方法结合。然后，通过计算，使用z的值选择信道，其中是连续向量。 对于离散设置，我们用one-hot编码**z**表示整数z，在这种情况下。 似然的最后一步是通过具有单个隐藏层的MLP传递h，然后是softmax，该softmax用于预测类标签y。

**2.2离散域标识符**

现在我们转到方法的中心部分，即先验的组件。 最简单的方法DSDA（请参见图1a）使用离散的潜变量，即是整数值的随机变量，因此可以将模型视为混合模型的一种形式。 该先验可预测给定输入x的z值，使用具有softmax输出的神经网络对其进行建模。 给定z，生成y的过程如2.1节所述。 可以使用目标为最大似然估计训练离散模型，



可以很容易地计算出，这是由（1）中的有限求和得出的，该求和要求分别计算k个分量中的每个分量，并对它们的结果求和。 该过程允许标准梯度反向传播，并以k线性缩放。

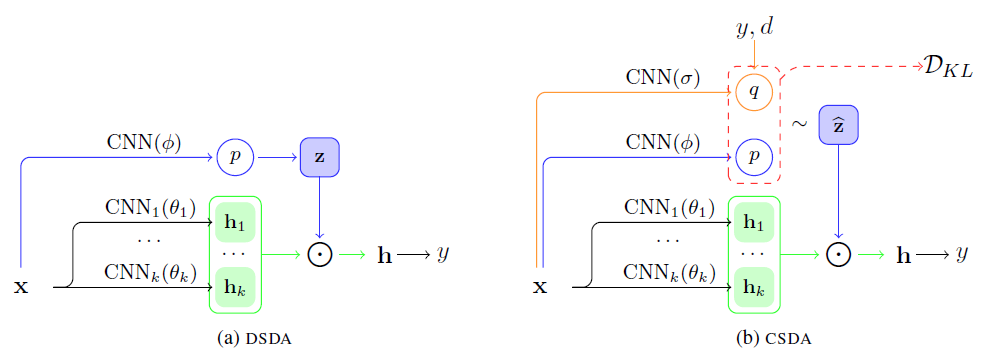


图1：潜在变量模型DSDA和CSDA的模型架构，在潜在变量的处理方面有所不同，这些变量是离散的（）或连续向量（）。 较低的绿色模型分量分别表示k个独立的卷积网络分量，蓝色和黄色分量分别表示先验p和变分近似q。 潜在变量用于选通k个隐藏的表示形式（显示为），然后将其用于线性函数中以预测分类标签y。 在训练期间，CSDA从q提取样本，而在推理期间，从p提取样本。

通过在观察到地面真实域d时将似然性最大化，可以将DSDA应用于有监督或半监督域。 我们将此设置称为“ DSDA + sup”。 或“ DSDA + semisup”，请注意，在此设置下，我们假设通道数k等于已知域D的清单。

**2.3连续域标识符**

为使DSDA模型正常工作，需要足够大的k，以便可以将所有不同类型的数据清楚地分为单独的混合成分。 如果在域之间没有清晰的界限，则推断的域后验可能不确定，该方法简化为合奏技术。 因此，我们引入第二种建模方法作为连续域标识符（CSDA），这受到LDA将文档建模为多个主题混合的方式的启发（Blei等，2003）。

在统计上更有效的方法是使用二进制函数作为域说明符，即，有效地实现了指数级的多个域组合（）。域zi的每个元素都充当门或等效地注意，控制是否将隐藏状态合并到预测模型中。 通过这种方式，模型的各个组成部分可以专门研究非常具体的主题，例如政治或体育，但是领域仍然能够将两者结合起来以产生专门的表示形式，例如体育政治。 潜在位向量的使用使推理变得棘手，这是由于在许多状态指数化下的边缘化。 因此，我们改为进行连续松弛，以使的每个标量zi从根据输入x进行参数化的概率分布中得出。 这些功能可以学习将x的各个方面与某些领域索引相关联，例如，使用诸如棒球之类的特定单词和一局就涉及与“运动”相对应的领域，从而允许自动学习文本领域。

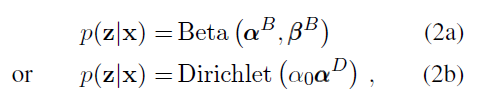
几种可能的分布可用于对建模。 在这里，我们考虑以下分布：

Beta，它将所有元素限制在[0; 1]，使得z位于超立方体中；

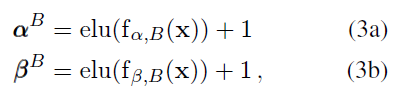
Dirichlet也限制了所有元素，如Beta一样，但是z也被约束在单纯性概率中。

在这两种情况下，z的每个维都由不同的分布参数控制，它们本身被公式化为x的不同非线性函数。基于Dirichlet模型在主题模型中的广泛使用以及其生成归一化向量的理想特性（类似于常见的注意力机制），我们期望Dirichlet模型表现最佳（Bahdanau等人，2015）。

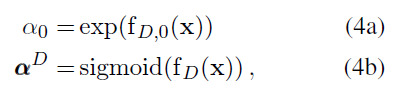
根据分布的选择，先验建模为



其中先验参数被参数化为输入的神经网络。 对于Beta版，



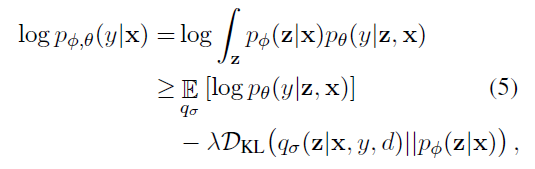
其中elu（）+ 1是一个元素激活函数，返回正值（Clevert等，2016），而是带有参数的非线性函数，此处我们使用CNN。 Dirichlet先验使用不同的参数化，



其中是一个正值总浓度参数，用于缩放（2b）中的所有成分，从而捕获整体稀疏度，而为与每个通道的亲和力建模。

**2.4变分推论**

如第2.3节所述，使用连续的潜在变量会产生棘手的推论。 因此，我们开发了基于变分自动编码器的变分推理方法（Kingma and Welling，2014）。 拟合模型涉及最大化证据下界（ELBO），



其中是通过进行参数化选择的变量分布，以匹配先验族（Beta或Dirichlet），并且是控制KL项权重的超参数。（5）中的ELBO相对于被最大化。 使用随机梯度上升，其中期望项使用单个样本近似，该样本用于直接计算似然度。尽管通常不可能通过样本反向传播梯度，这是学习变化参数所必需的，但通常使用重新参数化技巧来避开此问题（Kingma和Welling，2014年）。然而，该方法仅适用于有限范围的分布，尤其是高斯分布，因此，我们使用隐式重新参数化梯度方法（Figurnov等，2018），该方法可以推断各种连续分布，包括 Beta和Dirichlet。 我们在附录A.2中提供了有关隐式重新参数化方法的更多详细信息。

变异分布q以与先验p相似的方式定义，请参见（2-4b），即使用神经网络参数化来分配参数。 关键区别在于q不仅在x上而且在目标标签y和域d上都有条件。 这是通过嵌入y和d来完成的，将它们与x的CNN编码串联在一起，然后转换为分布参数。 通过在观察到域标识符时将d设置为域标识符，可以轻松地促进相对于域的半监督学习，对于域不受监督的实例，可以使用前哨值d = UNK。y使用相同的技巧，以允许进行原始半监督学习（关于目标标签）。y和d的使用允许推理网络学习将这两个关键变量编码为z，以鼓励潜在变量，从而为通道建模，从而为目标标签和域提供信息。 与（5）中的KL项相一致，这确保了先验p也必须学会仅基于输入文本x来区分域和标签。

为了在测试时进行推断，我们假设只有x可用作输入，因此无法使用推断网络。 取而代之的是，我们从先前的生成一个样本，然后将其用于计算最大似然标签。为了减少采样方差，我们还使用蒙特卡洛方法进行了试验推断，使用：（a）先验均值 （b）使用m = 100个来自先前样本的蒙特卡洛平均； （c）重要抽样（Glynn和Iglehart，1989年），根据来自推理网络q的抽样估计。尽管蒙特卡洛方法在10次运行中确实显示出很小的方差降低，但它们都没有显示出与单一样本技术相比在预测性能上的显着差异。 尽管它们的速度要慢几个数量级，所以此后我们将使用单个样本进行测试推断。

**3实验**

**3.1多域情感分析**

为了评估提出的模型，我们首先使用多域情感分析数据集进行实验，重点是测试域未知的域外评估。

我们从多域情感数据集v2.0（Blitzer et al。，2007）中获得数据集.5任务是预测二进制情感标签，即正值与否定值。 未处理的数据集具有20多个域。 为了我们的目的，我们筛选出标签实例少于1k的域或少于2k个未标记的实例，总共有13个域。

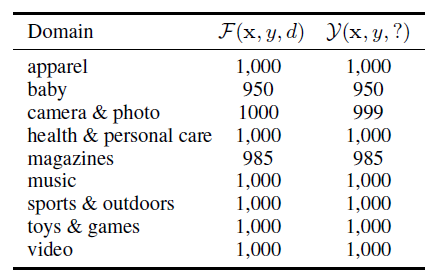


表1：我们数据集中每个训练域在F（域和标签已知）和Y（标签已知；域未知）两个类别下的实例（评论）数，其中“？” 表示“ UNK”令牌，表示给定的属性是不可观察的。

为了模拟半监督域情况，我们删除了一半标记数据的域属性，将它们表示为域未标记数据Y（x; y;？）。 另一半是情感和域标记的数据F（x; y; d）。我们在表1中列出了数据集的细分

为了进行评估，我们提供了四个领域，即书籍（“ B”），DVD（“ D”），电子产品（“ E”），厨房和家庭用品（“ K”），以便与以前的工作进行比较（Blitzer等 （2007年）。每个域都有1k个测试实例，我们将这些数据分成4：6的比例进行开发和测试。 dev数据集用于超参数调整和提前停止，我们报告测试的准确性结果。

**3.1.1基线和比较**

为了进行比较，我们使用3个基线。 第一个是单通道CNN（“ S-CNN”），它在单个模型中联合所有数据实例，而没有特定于域的参数。 第二个基线是多通道CNN（“ M-CNN”），它扩展了S-CNN模型（606k参数）的容量以匹配CSDA和DSDA（大约7.5m-8.3m参数）。 我们的第三个基准是使用对抗学习进行领域生成（GEN）的多领域学习方法，这是Li等人的最佳表现模型（2018a）和在可比较数据集上实现无监督多域自适应的最新技术。我们报告的结果他们表现最好的GEN + d + g模型。

**3.1.2训练策略**

对于超参数设置，我们在附录A.1中提供了详细信息。 在培训方面，如上所述，我们使用两种实验配置模拟了两种情况：（a）域监督； （2）域半监督。 对于域监督训练，仅使用F，它仅覆盖9个域，并且完全看不到测试域数据。 对于域半监督训练，我们使用F和Y的组合，请注意，两个子语料库均不包含来自目标域的数据，并且都没有明确地用情感y和域d标记。 它们模拟了我们具有包含大量相关数据的异构数据的设置，但是其元数据不一致，因此不容易建模。

对于（5）中的情况，根据ELBO的推导，应该是= 1，但是其他设置通常在实践中是合理的（Alemi等人，2018）。 因此，我们尝试了退火和固定时间表，但发现最终性能没有一致的差异。 我们对固定值= 进行了网格搜索；，根据开发性能选择=。 我们将在第3.2节中以敏感度图的形式提供进一步的分析。 DSDA的潜在域大小k设置为训练域的真实数量k = D =9。请注意，即使对于DSDA，我们也可以使用k / = D，我们将在第3.1.3节的F + Y监管设置中进行探讨。对于CSDA，我们给出了k = 13的主要结果，该结果与培训和测试中的域总数相匹配。

**3.1.3结果**

表2报告了两种训练配置下不同模型的性能：（1）具有F + Y（领域半监督学习）； （2）仅使用F（领域监督学习）。 在每种情况下，我们报告的标准差都是基于具有不同随机种子的10次运行。

总体而言，与先前的工作一致，域B和D比E和K困难。比较这两种配置，我们可以看到，当我们使用域半监督训练（加上Y）时，所有模型都将更好,演示了在带注释数据受限制的情况下领域半监督学习的实用性。

通过比较我们的离散方法和连续方法（分别是DSDA和DSDA），我们发现CSDA始终表现最佳，大大超过了基线。 相比之下，DSDA令人失望，其性能不如基准，而且在域监管与半监管或无监管设置之间，性能没有任何变化。 在基于CSDA的方法中，所有分布都表现良好，但Dirichlet分布总体上表现最佳，这归因于对域稀疏性的更好建模，从而减少了不确定域和混合域的影响。 对于领域半监督学习（F + Y）而言，最好的结果是，在不同类型的模型中，与领域监督学习（F）相比，其准确性提高了约2％。

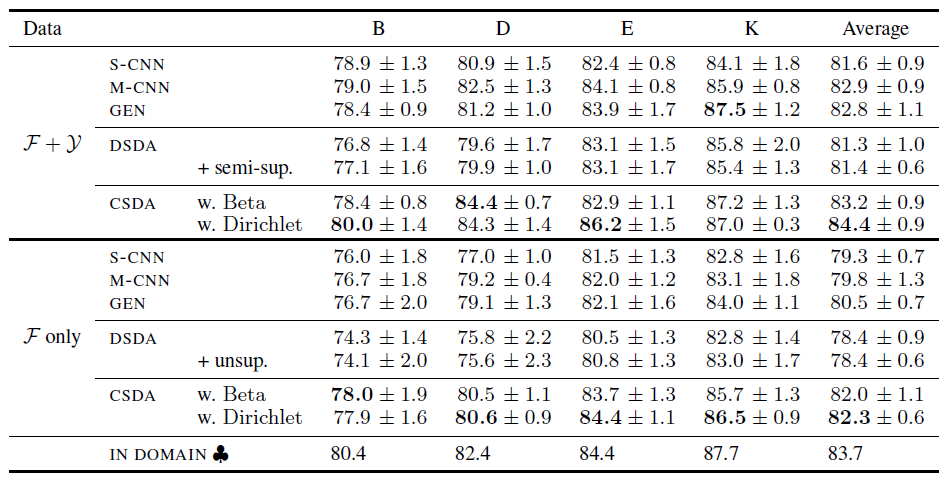


表2：在两种数据配置下，不同模型的准确性[％]和标准偏差：（1）同时使用F和Y（域半监督学习）； （2）仅使用F（领域监督学习）。 在每种情况下，我们都会评估四个保持不变的测试域（B，D，E和K），并报告准确性。 最佳结果在每种配置中均以粗体显示。来自Blitzer等。 （2007）。

**3.2分析和讨论**

为了更好地理解模型学习的内容，我们使用Dirichlet分布关注CSDA模型。

首先，我们根据潜在域大小k考虑模型容量。 图2显示了变化k的影响。 请注意，真实的域数是D = 13，包括9个训练域和4个测试域。 将k大致设置为该值似乎是合理的，因为平均精度随着k的增加而增加，并且在k = 16附近的平稳期。有趣的是，当k >32时，带有Beta的CSDA的性能下降，而Dirichlet的性能仍然很高-实际上，即使在k = 2的极值下，Dirichlet也始终是优越的，尽管随着k的增加它确实显示出改善。 还要注意，DSDA需要大量的潜在状态清单，这支持了我们关于连续离散的潜在变量cf效率的观点。

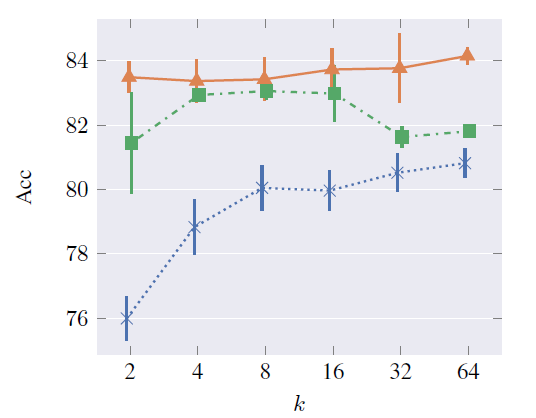
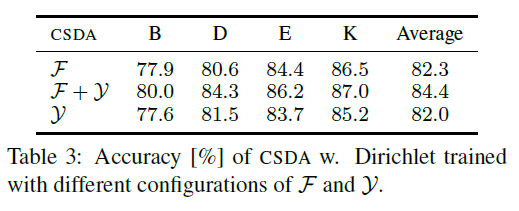


图2：使用DSDA（）和使用Beta（）和Dirichlet（）的三种CSDA方法在log 2空间中将潜在域大小k增加到F + Y时，具有标准误差（j）的性能。

接下来，我们考虑使用F和Y的不同组合的影响。表3显示了不同配置的性能。 总体而言，F + Y具有出色的性能。 有趣的是,Y本身仅比F差一点，表明目标标签y对于学习比域d更重要。 Y配置全域无监督训练仍然可以提供不错的性能，这非常适合应用于没有域元数据的非常混乱且异构的数据集。



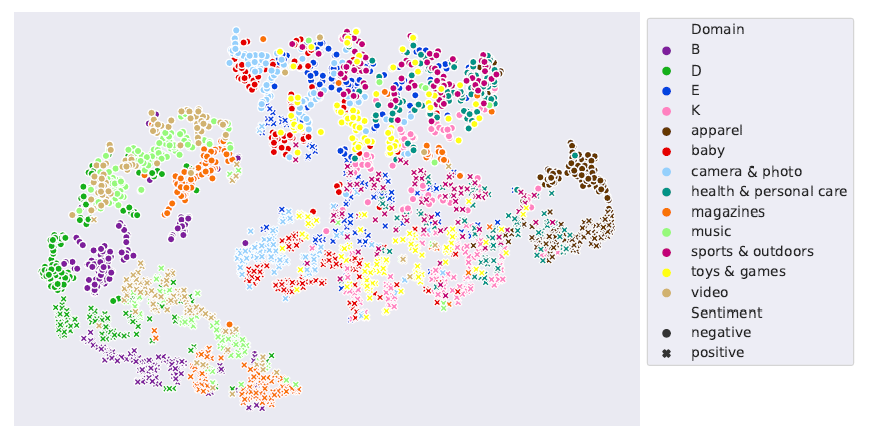


图3：CSDA在所有13个域中的隐藏表示的t-SNE，包括4个测试域（B，D，E和K），其余9个域仅用于训练。 每个点都是文档，符号表示其黄金情感标签，负数实例使用实心圆，正数使用十字形。

最后，我们将根据模型如何学习将k维潜在变量用于不同类型的数据来考虑模型正在学习的内容。 我们将学习的表示形式可视化，显示图3中2d tSNE图（Maaten和Hinton，2008）中绘制的每个域的点。请注意，每个域都分为两个簇，分别代表该域内的正（X）和负（）实例。 在测试领域中，B（书籍）和D（DVD）聚集在一起，但仍然清晰地分开，考虑到这两种媒体之间的紧密联系，这令人鼓舞。 另外两个（E（电子）和K（厨房和家庭用品））混合在一起并与其他领域混合在一起。 总的来说，在所有领域中，APPAREL集群都是截然不同的，VIDEO和MUSIC与D高度相关，而MAGAZINES的部分群集接近B； 考虑到各个产品之间的相似性，所有这些都直观直观。E与CAMERA和GAMES有关，而K与HEALTH和SPORTS关系最密切。

为了更好地了解潜在变量中编码的内容，以及如何通过设置来实现这一点，我们学习了简单的诊断分类器来预测情感标签y和域标签d（仅将z作为输入）。 为此，我们首先在训练集上训练模型，然后记录来自推理网络的z样本。然后，我们对训练集进行划分，使用70％的学习线性逻辑回归分类器来预测y和d，然后使用剩余的30％进行评估。

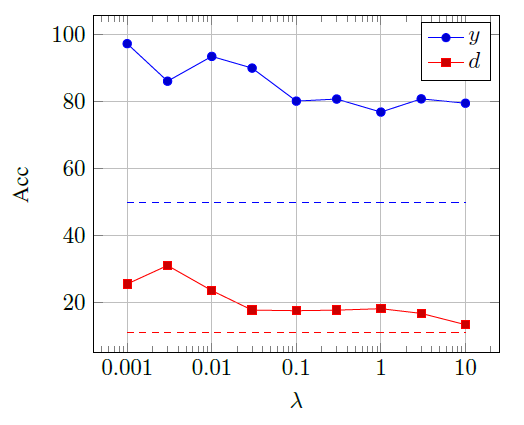


图4：诊断分类器相对于z的准确度[％]，以对数刻度显示相对于不同的的情感标签y和域标签d。 虚线水平线表示两个输出的机会准确性。

图4显示了基于三个运行的平均值的预测准确性，每个运行具有不同的z个样本。 显然，很小的会导致几乎完美的情感标签准确性，这是通过使用潜变量对响应变量进行编码而过度拟合的证据。 正如预期的那样，对于而言，情绪准确性仍然高于偶然性，但更为稳定。 对于域标签d，预测准确性也较高，尽管程度较小，并且显示出相似的下降趋势。 在较早的实验中使用的设置为时，这表明潜变量编码捕获了实质性的情感和一些领域知识，如图3所示。

就训练所需的时间而言，使用默认设置的CSDA方法的训练时间约为25分钟，而DSDA和M-CNN的训练时间相似。 运行时间随着潜在大小k的增加而呈亚线性增加。

3.3语言识别

为了进一步证明我们的方法，我们然后在第二个任务语言识别中评估我们的模型（LangID：Jauhiainen等（2018））。对于数据处理，我们遵循Lui和Baldwin（2011）的设置，使用来自5个不同领域的5种训练集，使用97种语言。 我们评估了7个支持基准的准确性：来自Baldwin和Lui的EUROGOV，TCL，WIKIPEDIA（2010），EMEA（Tiedemann，2009），EUROPARL（Koehn，2005），TBE（Tromp和Pechenizkiy，2011）和TSC（Carter等，2013）。 与情感任务不同，在这里，我们使用完整的数据集评估我们的方法，但有两种配置：（1）域不受监督，其中所有实例都只有标签而没有域（表示为Y）； （2）领域监督学习，其中所有实例都有标签和领域（F）。

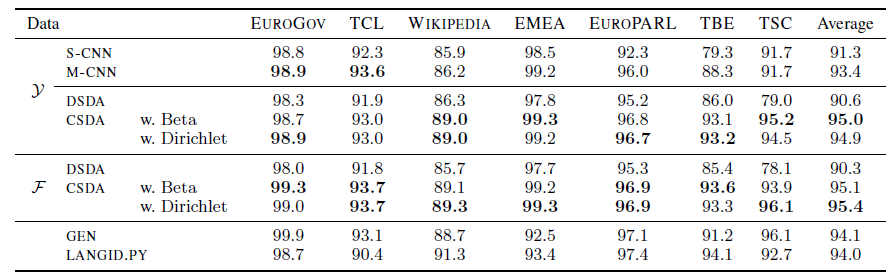


表4：在两种数据配置下，不同模型在7个LangID基准测试中的准确度[％]以及平均分数：（1）使用域无监督学习（Y）； （2）使用领域监督学习（F）。 在每种配置中，最好的结果以粗体显示。 请注意，GEN和LANGID.PY的训练数据与原始论文中使用的训练数据略有不同。

**3.3.1结果**

表4显示了不同模型在7个保持基准上的表现以及平均分数。 我们还报告了GEN（Li等人的最佳模型）的结果。 （2018a），以及一个最新的现成的LangID工具：LANGID.PY（Lui和Baldwin，2012年）。 注意，S-CNN和M-CNN都是域无监督方法。 就结果而言，总体而言，我们的两个CSDA模型始终优于所有其他基线模型。

比较不同的CSDA变体（Beta和Dirichlet），它们在LangID任务中的执行情况都非常接近。 此外，CSDA在平均分数方面的表现优于最新技术。 有趣的是，这两种训练配置表明，领域知识F对CSDA的性能提升不大，但对DSDA却无济于事。 最重要的是，LangID结果证实了我们提出的方法的有效性。

4相关工作

领域适应（“ DA”）通常涉及一个或多个培训领域和一个目标领域。在DA方法中，单域适应是最常见的方案，其中模型在一个域上训练，然后使用先验对目标域的了解（Blitzer等人，2007； Glorot等人，2011）模型转移到单个目标域。已经提出了对抗性学习方法来学习鲁棒的领域无关表示，可以通过半监督学习来捕获领域知识（Ganin et al。，2016）。

多域适应使用来自多个训练域的训练数据。 方法包括特征增强方法（DauméeIII，2007）和类似的神经模型（Joshi等，2012； Kim等，2016），以及基于注意力和层次的方法（Li等，2018b）。这些工作假设在传输时已知“ oracle”源域，但是在本文中我们不需要oracle。 对抗训练方法已被用来学习鲁棒的领域广义表示（Liu et al。，2016）。 Li等。 （2018a）考虑了模型无法访问目标域的情况，并使用对抗学习通过源域之间的交叉比较来生成域生成表示。

这项工作的另一个重要组成部分是变分推理（VI），这是一种来自机器学习的方法，可以通过优化来近似概率密度（Blei等，2017； Kucukelbir等，2017）。 变体自动编码器的思想已应用于语言生成（Bowman等人，2016; Kim等人，2018; Miao等人，2017; Zhou和Neubig，2017; Zhang等人，2016）和 机器翻译（Shah and Barber，2018; Eikema and Aziz，2018），但不是在半监督领域适应的背景下。

**5总结**

在本文中，我们提出了两种用于多域学习的模型DSDA和CSDA，它们使用带有潜在变量的图形模型来表示域。 我们提出了具有离散潜变量和连续向量值潜变量的模型，我们使用Beta或Dirichlet先验模型进行建模。 对于训练，我们采用基于变分自动编码器的变分推理技术。 在对多域情感数据集和七个语言识别基准进行的经验评估中，我们的模型在各种数据条件（包括未提供目标域数据的设置）下均优于强基准。 我们提出的模型在异构语料库的NLP应用中具有广泛的实用性。