

文章编号:

基于深度表示学习和高斯过程迁移学习的情感分析方法*

吴冬茵¹, 桂林¹, 陈钊², 徐睿峰^{1*}

(1. 哈尔滨工业大学 深圳研究生院计算机科学与技术学院, 广东 深圳 518055;

2. 腾讯科技(深圳)有限公司, 广东 深圳 518055)

摘要: 情感分析是自然语言处理领域的重要研究问题。现有方法往往难以克服样本偏置与领域依赖问题, 严重制约了情感分析和应用。为此, 本文提出了一种基于深度表示学习和高斯过程知识迁移学习的情感分析方法。该方法首先利用深度神经网络获得文本样本的分布式表示, 而后基于深度高斯过程, 从辅助数据中迁移与测试集数据分布相符的高质量样例扩充训练数据集用于分类器训练, 以此提高文本情感分类系统性能。在 COAE2014 文本情感分类数据集上进行的实验结果显示本文提出方法性能可以有效提高文本情感分类性能, 同时可以有效缓解训练数据的样本偏置以及领域依赖问题的影响。

关键词: 情感分析; 深度表示学习; 高斯过程; 迁移学习

中图分类号: TP391

文献标识码: A

Sentiment Analysis based on Deep Representation Learning and

Gaussian Processes Transfer Learning

Dongyin Wu¹, Lin Gui¹, Zhao Chen², Ruifeng Xu^{1*}

(1. School of Computer Science and Technology, Harbin Institute of Technology Shenzhen

Graduate School, Shenzhen, Guangdong, 518055, China;

2. Tencent Technology (Shenzhen) Ltd., Shenzhen, Guangdong, 518055, China)

Abstract: Sentiment analysis is an important topic in natural language processing research. Most existing sentiment analysis techniques are difficult to handle the domain dependent and sample data bias problems, which puzzles the development and application of sentiment analysis. Target to these problems, this paper presents a sentiment analysis approach based on deep representation learning and Gaussian Processes transfer learning. Firstly, the distributed representations of text samples are learned based on deep neural network representation learning. Next, based on deep Gaussian processes, this approach selects quality samples with the distribution similar to testing dataset from additional dataset to expand the training dataset. The sentiment classifier trained on the expanded dataset is expected to achieve higher performance. The experimental results on COAE2014 dataset show that the proposed approach improved the sentiment classification performance. Meanwhile, this approach is helpful to reduce the influences of training sample bias and domain dependence.

Keywords: Sentiment Analysis; Deep Representation Learning; Gaussian Processes; Transfer Learning

1 引言

互联网和移动互联网技术的飞速发展,吸引了大量用户在网络平台上分享生活,表达观点。随着海量用户生成文本的不断积累,对文本中包含的情感进行分析,获取用户对特定事件、人物或产品的评价与态度的文本情感分析研究正在成为自然语言处理领域的研究热点。

传统的文本情感分析技术主要分为基于规则的方法和基于统计的机器学习方法。前者依

* 收稿日期: 定稿日期:

基金项目: 国家自然科学基金(61370165),国家 863 计划(2015AA015405), 深圳市孔雀计划技术创新项目(KQCX20140521144507925), 深圳市基础研究项目(JCYJ20150625142543470), 广东省数据科学工程技术研究中心开放课题(2016KF09)

托语言学研究成果, 主要利用人工构建规则和情感词典, 对文本的情感倾向性进行分析^[1]。后者主要利用人工标注的训练语料, 提取文本特征并建立统计模型, 实现文本情感倾向性的分类。近年来, 基于深度表示学习的方法在文本情感分析任务中也得到很大发展。目前主流的基于有监督学习的方法需要大量人工标注的训练语料来提高模型的学习能力和泛化能力, 但人工标注代价往往偏高, 面向实际应用的大量标注数据也很难获得。另外, 基于统计机器学习的方法通常假设训练语料和测试语料是独立同分布的, 然而这个假设在实际应用中往往难以得到满足。这就意味着这些方法难以克服样本偏置和领域依赖的问题。

针对这些问题, 本文提出了一种基于深度表示学习和高斯迁移过程的文本情感分析方法。该方法首先利用词向量模型和卷积神经网络获得文本样本的分布式表示。而后, 基于 k 最近邻的深度高斯过程设计了一种非对称迁移学习方法, 通过从辅助数据中迁移与测试集数据分布相符的高质量样例扩充训练数据集, 实现迁移学习用于分类器训练。由于这一方法不需要满足独立同分布, 可以使用从一个领域学习到的知识帮助学习新领域的知识, 因此有望提高文本情感分类的性能。在 2014 中文观点倾向性分析评测 (Chinese Opinion Analysis Evaluation, COAE2014) 数据集上的实验显示, 本文提出的方法性能优于目前主流的卷积神经网络 (Convolutional Neural Network, CNN^[2]) 和基线的朴素贝叶斯支持向量机 (Naïve Bayes Support Vector Machines, NBSVM^[3]) 方法, 显示出本文提出的思路可以有效提高文本情感分类性能, 同时可以缓解训练数据样本偏置以及领域依赖的影响。

本文的内容组织如下: 第二节简单回顾情感分析以及迁移学习的相关研究工作; 第三节介绍本文提出的基于深度表示学习和高斯过程迁移学习的情感分析方法; 第四节通过实验评估本文提出的方法的有效性; 第五节给出本文的总结和展望。

2 相关工作

文本情感分析研究旨在通过分析文本的语义倾向性特征, 识别出文本所包含的情感倾向。目前, 文本情感分析技术大体分为基于规则的方法和基于统计机器学习的方法。前者主要利用语言学知识, 如语言表达规则和情感词典进行分析。Taboada^[4]利用情感词典及词性定义情感词的语义倾向得分, 并将其按特定规则组合计算, 以此判断文本的情感倾向性。Fu^[5]等根据模糊集合和量化情感词典解决情感分类及词语极性强烈程度分级。Bollegala^[6]等人利用不同领域情感表达方式的共性来构建领域相关的情感词典, 扩充文本特征, 以此提高跨领域情感分类的效果。基于规则的方法可以直观地给出文本情感倾向性的计算方法, 但其性能受限于人工规则的完备程度和情感词典的覆盖率限制, 难以快速处理新类型的文本数据。

基于统计机器学习的情感分析方法利用人工标注的训练语料, 提取文本情感倾向性特征并建立统计模型, 从而自动判别未知数据的文本情感倾向性。Pang^[7]采用基于 N-gram 的词语特征和词性特征, 将朴素贝叶斯 (Naive Bayesian, NB)、最大熵 (Maximum Entropy, ME) 和支持向量机 (Support Vector Machines, SVMs) 三个机器学习模型应用到文本情感分类, 为后续的研究提供了参考。Zhao^[8]提出利用句子上下文信息和添加冗余标签来建立基于条件随机场 (Conditional Random Fields, CRFs) 的三层分类模型, 有效降低了分类错误在不同层级间的传播。近几年, 基于深度神经网络和表示学习的方法在文本情感分析研究中得到了较多应用。Socher^[9]在句法分析树的基础上, 利用循环神经网络对电影评论文本进行情感倾向性分类, 分别在细粒度情感分析数据集和正负情感倾向性分析数据集上取得了显著的效果。Kim^[2]利用 word2vec 训练得到的词向量, 将文本数据映射到文本特征矩阵, 利用卷积神经网络实现对文本的情感倾向性分类。Sun^[10]使用微博评论扩展原微博文本, 结合多层限制玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 与深度神经网络模型对短文本进行情感分析。基于统计的机器学习方法, 其学习能力和泛化能力的提升需要大量人工标注的训练语料, 这

在实际应用中往往难以获得,导致其性能受到影响。另一个问题是训练数据和测试数据难以满足独立同分布假设。对于训练数据和测试数据来自不同领域,或来自同一领域但训练数据无法覆盖整个特征空间的情况,从训练数据中得到的模型有偏差,无法准确预测未知数据。

考虑到迁移学习不需要假设训练数据和测试数据符合独立同分布的要求,而是用从一个领域学习到的知识来帮助学习新领域的知识。通过迁移学习,可以将从已有数据学习到的知识迁移应用到新的领域来弥补新领域训练数据不足的缺陷。因此,迁移学习方法得到了广泛重视。目前大部分针对迁移学习的研究主要集中在什么知识可以迁移以及如何迁移。迁移学习大致可以分为三类:归纳式迁移学习、转换式迁移学习和无监督迁移学习^[11]。归纳式迁移学习从训练数据提取适合测试数据的样例或特征进行知识迁移,如从训练数据中抽取与测试数据相似的训练实例或者提取两者共有的特征。Dai^[12]假设训练数据和测试数据在同一特征空间中,然后基于提升方法的思想不断调整训练数据中有效样本的权重,同时降低噪音数据的影响。Argyriou^[13]则提出针对多任务迁移学习的稀疏特征学习算法以解决在多任务之间的迁移学习问题。转换式迁移学习要求模型训练期间能够利用来自不同领域测试数据。受到重要性采样思想的启发,转换式迁移学习集中研究如何估计训练数据中每个样例的权重,以此实现经验风险最小化。Zadrozny^[14]通过定义简单分类问题来估计训练数据和测试数据的先验分布比值,以此衡量每个训练样本的权重。Fan^[15]则进一步分析了多种分类器对估计训练数据和测试数据的先验分布比值的实现效果。无监督迁移学习则在没有标注数据的前提下,从训练数据和测试数据中学习能够表示训练数据和测试数据的共有特征。

在文本情感分析领域可以应用迁移学习解决单一领域人工标注数据不足、样本选择偏置、跨语言情感分析等问题。考虑到文本数据具有领域相关性,不同领域的情感表达具有不同的方式,主要体现在相同的情感词、评价词在不同领域可能表达不同的情感倾向性,所以在利用不同领域的数据进行迁移学习时如何选择合适的特征和样本进行迁移成为研究的重点。Blitzer^[16]利用少量标注的目标领域数据来修正结构对应学习方法(Structural Correspondence Learning, SCL)在源数据和目标数据上的不重合问题,并采用无监督的样本相似度度量方法来从源数据中选择最佳样本。Xu^[17]在处理跨语言文本情感分析的时候,提出基于实例的迁移学习,将大量含有标注的源语料翻译成目标语言,并从中筛选出有效样例来扩充目标语料数据集。Gui^{[18][19]}则通过检测迁移学习中出现的噪音数据来降低迁移学习带来的负面影响,从而获得跨语言情感分析的效果提升。

3 基于深度表示学习和高斯过程迁移学习的文本情感分析方法

3.1 基于词向量和卷积神经网络的文本深度表示学习

传统的基于词袋模型的文本表示存在维度过高、数据稀疏、分布不连续的缺陷,相对不适合对文本样本进行表示,本文应用了基于词向量和卷积神经网络的文本深度表示学习方法。

卷积神经网络模型是一种有监督学习的模型,根据输入数据类型的不同,可以由不同数量的卷积层和下采样层叠加组成,并最终将特征输出到全连接层。卷积层可以提取输入数据的抽象特征,包含多个神经元,这些神经元之间共享权值,从而减少了模型的参数。下采样层属于特征映射层,由多个卷积层产生的特征图组成,通过对固定大小的区域上不同位置的特征进行聚合计算,如平均值计算或者最大值计算,输出局部区域内的最优特征。经过下采样的池化处理,不仅降低了卷积特征层的特征维度,而且还保留了最有效的特征。全连接层由普通神经网络构成,利用反向传播算法实现模型参数的更新。

卷积神经网络模型的输入数据为二维特征矩阵,因此像图像等二维数据可以直接作为卷积神经网络模型的输入。而文本数据往往被表示为一维特征向量,需要将其扩展为二维特征

矩阵的表示形式。以 word2vec 为代表的分布式词向量的出现，为文本数据映射到二维特征矩阵提供了可行的方案。本文借鉴 Kim^[21]提出的基于 word2vec 训练的词向量的 CNN 模型(下文标记为 W2VCNN)，将文本数据映射到序列特征。以 word2vec 模型的分布式词向量为输入的文本数据，将文本数据转化为特征向量，实现了文本的分布式表示学习。利用上述特征，可以分类器进行特征学习，构建分类模型。

3.2 基于深度高斯过程的非对称迁移学习

高斯过程又被称作正态随机过程，满足正态分布，属于自然界普遍存在的一种随机过程。高斯过程在处理高维度、非线性等复杂问题上，不仅具有良好的适应性，而且具有较强的模型泛化能力。在分类预测模型上，可以将高斯过程回归模型的输出映射到一个概率值上，表示对应输出类别的概率。Kandemir^[20]提出了一种基于深度高斯过程的非对称迁移学习模型 (Asymmetric Transfer Learning with Deep Gaussian Processes, ATL-DGP)，并在图像分类领域取得了应用。图 1 为 Kandemir 提出的 ATL-DGP 模型的示意图。本文首先在该模型的基础进行扩展，实现该模型在文本情感分析上的应用。

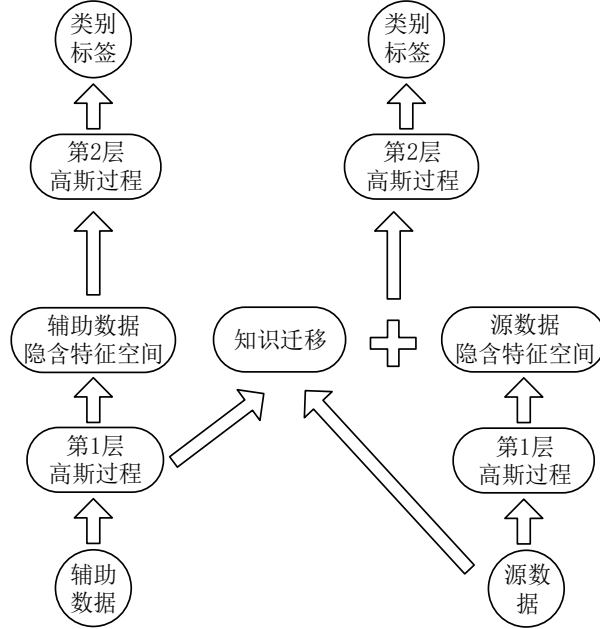


图 1 基于深度高斯过程的非对称迁移学习模型

设源数据集为 $D_{source} = \{X_s, Y_s\}$ ，辅助数据集为 $D_{extra} = \{X_e, Y_e\}$ ，目标数据集为 $D_{target} = \{X_t, Y_t\}$ 。分别以源数据集 D_{source} 和辅助数据集 D_{extra} 为训练数据，建立两个高斯过程分类模型 $F = \{F_s, F_e\}$ 。

根据高斯过程模型的定义，对于高斯过程分类模型 F_e 有：

$$p(Y_e|F_e) = \mathcal{N}(Y_e|F_e, \beta^{-1}I) \quad (1)$$

$$p(F_e|D_e) = \mathcal{N}(F_e|0, K_{D_e D_e}) \quad (2)$$

其中 $p(\cdot)$ 表示密度函数， $\mathcal{N}(x|\mu, \Sigma)$ 表示具有均值 μ 和协方差 Σ 的高斯分布。式(1)和下文中 α 、 β 和 λ 为高斯分布的参数。 B_e 为辅助数据 X_e 在高斯过程分类模型 F_e 的第一层高斯过程中产生的非线性特征映射， D_e 表示 B_e 在高斯过程分类模型 F_e 的隐含特征空间的表示。 $K_{D_e D_e}$ 为高斯过程分类模型 F_e 在第二层高斯过程模型中基于隐含特征向量 D_e 产生的协方差矩阵。那么，根据高斯过程分类模型 F_e 的第一层高斯过程，可以得到：

$$p(D_e|B_e) = \prod_{r=1}^R \mathcal{N}(d_e^r|b_e^r, \lambda^{-1}I) \quad (3)$$

$$p(B_e|X_e) = \prod_{r=1}^R \mathcal{N}(b_e^r|0, K_{X_e X_e}) \quad (4)$$

$K_{X_e X_e}$ 是基于第一层高斯过程的核函数在辅助数据 X_e 生成的协方差矩阵。 R 为隐含节点数量。

利用源数据集 $\{X_s, Y_s\}$ 可以构建类似的高斯过程分类模型 F_s :

$$p(Y_s|F_s) = \mathcal{N}(Y_s|F_s, \beta^{-1}I) \quad (5)$$

$$p(F_s|D_s) = \mathcal{N}(F_s|0, K_{D_s D_s}) \quad (6)$$

$$p(D_s|B_s, B_{e \rightarrow s}, \pi) = \prod_{n=1}^{N_t} \mathcal{N}(d_s^n | \pi b_{e \rightarrow s}^n + (1 - \pi) b_s^n, \alpha^{-1}I) \quad (7)$$

$$p(\pi) = \text{Beta}(\pi|e, f) \quad (8)$$

$$P([B_{e \rightarrow s} \oplus B_s] | X_e, X_s) = \prod_r^R \mathcal{N}([b_e^r \oplus b_s^r] | 0, K_{[X_e \oplus X_s][X_e \oplus X_s]}) \quad (9)$$

$$p(B_s|X_s) = \prod_{r=1}^R \mathcal{N}(b_s^r | 0, K_{X_s X_s}) \quad (10)$$

根据上述推导公式，可以发现，高斯过程分类模型 F_s 与高斯过程分类模型 F_e 不同在于隐含特征空间 D_s 的计算。在高斯过程分类模型 F_s 中， $B_{e \rightarrow s}$ 由模型 F_e 的第一层高斯过程对源数据进行非线性映射产生。模型 F_s 的隐含特征表示 D_s 由 $B_{e \rightarrow s}$ 和 B_s 进行线性加权组合产生。权重系数 π 服从 **Beta** 分布。上述过程实现了知识迁移，是整个基于深度高斯过程的非对称知识迁移学习的核心部分。

高斯过程模型在处理大规模训练数据时，求解协方差矩阵的逆将成为整个模型的计算瓶颈，基于深度高斯过程的非对称迁移学习模型在模型训练过程中同样会面临该问题。**Kandemir** 在 **ATL-DGP** 模型中采用了 **Snelson**^[21] 提出的基于伪输入数据的稀疏高斯过程求解方法。该方法假定输入数据满足独立同分布条件，因此可以通过对输入数据进行随机采样，获得输入数据的子集，并构建低阶矩阵来近似完全数据高斯过程计算的结果。

3.3 基于 k 近邻的深度高斯过程的非对称迁移学习

在文本的情感分析领域，由于训练数据有限，文本数据映射到特征空间后，有限的训练样本无法涵盖整个文本数据的特征空间，出现训练数据和测试数据分布不一致的现象，产生样本选择偏置的问题。在训练数据集上得到的最优分类器并不一定能够很好地预测未知数据。如图 2 所示，黑色图形代表训练数据集上的实例，白色图形代表测试数据集上的实例，斜线为以训练数据集为基础构建的分类器的分类决策面。

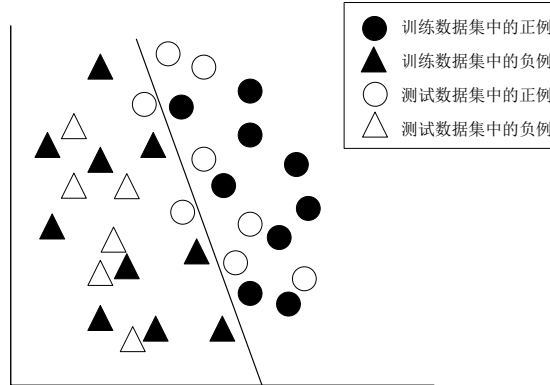


图 2 在有偏置训练数据集上的数据分类示意图

从图 2 中可以看出，以训练数据集为基础构建的最优分类器往往可以在训练数据集上取

得最好效果.但是当分类器应用到测试数据的时候,数据分布不一致带来的问题就凸显出来。因此,通过随机采样方法获取的数据子集的分布不一定符合测试数据的分布。如果先对训练数据进行预处理,筛选出符合测试数据分布的实例集,再从这个实例集里面去采样得到一个可以近似求解高斯过程的子集,就可以尽可能地降低噪音数据对模型影响,从而提高模型分类效果。基于上述假设,本文在 ATL-DGP 模型的基础上,设计一个基于 k 近邻 (k -Nearest Neighbor, k NN) 方法应用于伪数据子集的选择过程,从而将伪数据子集的选择范围限制在符合测试数据分布的训练实例上,保证了伪数据子集的质量,避免随机选择带来的负面影响。

对于离线学习问题,我们虽然不知道测试数据的具体类别标签,但是可以得到测试数据的分布信息。利用测试数据的分布信息,从训练数据中挑选符合测试数据分布的实例,并基于这些实例构建分类器,以此得到近似满足测试数据分布的最优分类器。因此,本文提出结合 k 近邻的方法,从训练样本中,挑选近似满足测试数据分布的样本构成新的训练数据集。具体过程见算法 1 所示。

算法 1. 基于 k 近邻的样本选择算法

输入: 训练样本集 D_{train} , 测试样本集 D_{test} , 文本特征映射模型, 近邻数 k ;

输出: 新的训练样本集 D_{new} 。

1. 根据文本特征映射模型将训练样本和测试样本映射为特征向量集
 2. 基于 D_{train} 构建 k NN 分类器 C_I
 3. 新训练数据集 D_{new}
 4. **For each instance** $v \in D_{test}$
 5. 根据 k NN 分类器 C_I 获取 v 的 k 近邻实例 $v_1 \dots v_k \in D_{train}$
 6. $D_{new} = D_{new} \cup \{v_1 \dots v_k\}$
 7. **End for**
-

图 3 示意了经过实例选择后训练得到的分类模型的最优分界面。虚线为样本选择前的最优分界面。可以看到经过 k 近邻样本选择后,与测试样本相似的训练样例得以保留,不相似的视为噪音被去除,对应地最优分界面移到实线位置,有望得到更好的分类效果。

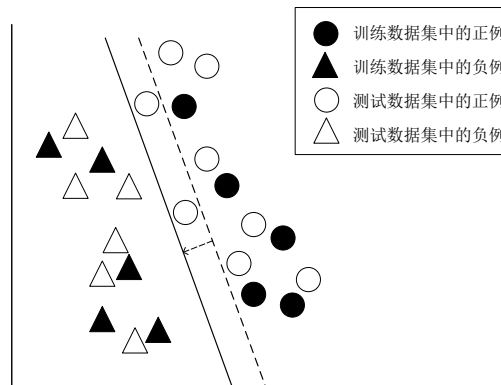


图 3 在相似数据集上的分类效果示意图

结合 ATL-DGP 模型,我们使用 k NN 样本选择算法对其扩展以消除部分样本选择偏置。扩展后的算法标记为基于 k 近邻的深度高斯过程的非对称迁移学习方法 (k -nearest Neighbors Asymmetric Transfer Learning with Deep Gaussian Processes, k ATL-DGP)。

算法 2. 基于 k 近邻的深度高斯过程的非对称迁移学习模型算法

输入：源数据集 D_{source} ，辅助数据集 D_{extra} ，目标数据集 D_{target} ，文本特征映射模型， k 近邻样本选择算法参数 k ，隐含节点个数 R ；

输出：目标数据集 D_{target} 的类别标签。

1. 利用源数据集 D_{source} 和辅助数据集 D_{extra} 训练得到文本特征映射模型 M
 2. 根据文本特征映射模型 M 将源数据集 D_{source} 、辅助数据集 D_{extra} 和目标数据集 D_{target} 映射为特征向量
 3. 利用 k 近邻样本选择算法分别从 D_{source} 和 D_{extra} 中得到伪数据集 P_{source} 和 P_{extra}
 4. 利用数据集 D_{source} 、 D_{extra} 、 P_{source} 、 P_{extra} 训练得到 ATL-DGP 模型
 5. 利用 ATL-DGP 模型预测目标数据集 D_{target} 并输出分类标签
-

4 实验结果及分析

为评估本文提出方法的性能，本节以 2014 中文观点倾向性分析评测（COAE2014）数据集中的训练数据集作为源数据集，测试数据集作为目标数据集，同时引入 2014 年自然语言处理与中文计算会议（Natural Language Processing & Chinese Computing, NLPCC2014）基于深度学习的情感分类评测的训练数据集作为辅助数据集。

COAE2014 微博数据集来源于第六届中文倾向性分析评测之微博观点句识别子任务，分为训练数据集和测试数据集。训练数据来自同一话题，总共 2174 条，其中有 1003 条带有正面情绪，1171 条带有负面情绪。测试数据分别选自手机、保险、翡翠三个不同话题，总共 7000 条，其中带有正面情感的有 3776 条，带有负面情感的有 3224 条。

NLPCC2014 情感分类评测数据集的数据来源于网络评论数据，涵盖数字影音、手机数码、图书等领域。其训练数据集包含 5000 条正面评论数据和 5000 条负面评论数据。

4.1 实验数据预处理与参数设置

在数据预处理方面，首先去除文本中数据中的噪音信息，如表情符号、电子邮件地址、用户昵称等。然后利用 ICTCLAS¹分词工具对实验数据集进行分词和词性标注。在文本特征映射过程中，使用 Google 开源的 word2vec²中 Skip-gram 模型训练 W2V CNN 的分布式词向量，以 2000 万条无标注的微博数据作为训练语料，生成包含 33 万个词汇的分布式词向量模型。对于分布式词向量模型中不包含的词语，采用随机生成分布式词向量的方式来产生未登录词语的分布式词向量。具体参数见表 1。

基于高斯过程的知识迁移学习分别为源数据和辅助数据构建一个两层的高斯过程分类模型，通过对层与层之间的隐含特征的线性组合实现辅助数据到源数据的知识迁移。其中 k ATL-DGP 模型采用表 2 所示的参数设置。

¹ <http://ictclas.nlp.ir.org/>

² <http://word2vec.googlecode.com/svn/trunk/>

表 1 W2VCNN 模型的参数设置

序号	参数类型	可调参数	值
1		模型训练算法	Negative Sampling
2	Skip-gram 模型参数	滑动窗口大小	5
3		词向量维度	50
4		学习率	0.025
5		卷积核函数	Rectified linear 函数
6	卷积神经 网络参数	过滤器滑动窗口大小 h	2,3,4,5
7		过滤器数量 m	100
8		随机更新参数比例	0.5
9		逻辑回归衰减系数	0.95
10		训练迭代次数	50
11		批处理数量	100

表 2 k ATL-DGP 模型参数设置

序号	可调参数	值
1	核函数	径向基核函数
2	伪数据集大小	10
3	源数据隐含节点个数	10
4	辅助数据隐含节点个数	10
5	学习率	0.001
6	k 近邻个数	1
7	训练迭代次数	1

4.2 实验结果及分析

本节分别评估了基于非迁移学习方法包括 NBSVM、W2VCNN₁、W2VCNN₂、RBMDNN，以及结合迁移学习方法 ATL-DGP、 k ATL-DGP 的性能。其中 NBSVM 由 Wang^[3]提出，采用 Unigram 和 Bigram 语言模型构建文本特征向量，并采用线性支持向量机作为模型分类器。RBMDNN^[10]方法结合了限制玻尔兹曼机和深度神经网络，在短文本分类上获得了很好的效果。实验中，NBSVM、W2VCNN₁ 仅使用 COAE2014 训练数据集作为训练数据，W2VCNN₂ 使用 COAE2014 训练数据集和 NLPCC2014 情感分类评测数据集训练 CNN 模型，迁移学习方法 ATL-DGP、 k ATL-DGP 均采用 COAE2014 训练数据集作为源训练数据集、NLPCC2014 情感分类评测数据集作为辅助数据集。实验所用测试数据集均为 COAE2014 测试数据集。

实验结果如表 3 所示，可以看到对比于非迁移学习方法 NBSVM、W2VCNN₁ 和 RBMDNN，结合了迁移学习的方法 ATL-DGP 和 k ATL-DGP 达到了更高的性能。此外，虽然迁移学习方法使用了辅助数据集，但同样加入了 NLPCC 训练集的 W2VCNN₂ 模型的实验结果表明，训练数据规模的增大并不是性能提高的主要原因。相反，W2VCNN₂ 对比 W2VCNN₁ 仅在负面情绪识别时 F 值上升 1.9%，而正面情绪识别性能 F 值下降 4.0%，说明由于辅助数据集和源训练数据集不同源，可能引入了噪音反而造成训练数据质量下降。在实验中，引入辅助数据的迁移学习方法 ATL-DGP 和 k ATL-DGP 表现出对样本选择偏置的较强的处理能力。

表 3 情感分类算法在 COAE2014 数据集上的性能对比

模型	Precision+	Recall+	F-Score+
	Precision-	Recall-	F-Score-
NBSVM	0.7175	0.8273	0.7685
	0.7536	0.6185	0.6794
W2VCNN ₁	0.7870	0.8864	0.8337
	0.8438	0.7190	0.7764
W2VCNN ₂	0.8658	0.7328	0.7937
	0.7348	0.8669	0.7954
RBMDNN [*]	0.8620	0.7300	0.7910
	0.7270	0.8900	0.8000
ATL-DGP	0.8613	0.8271	0.8438
	0.8065	0.8440	0.8248
<i>k</i> ATL-DGP	0.8680	0.8236	0.8452
	0.8051	0.8533	0.8285

(※所用实验设置与本文不同,采用 COAE2014 任务 4 部分训练数据作为训练集,其余数据为测试集。)

与 W2VCNN₁ 对比, ATL-DGP 在非同源的 NLPCC 数据集的辅助下, 迁移学习策略显示出一定成效,其正面情绪识别的 F 值提升 1%, 负面情绪识别的 F 值提升较大, 达到 4.84%。此外, 本文提出的 *k*ATL-DGP, 相较于 ATL-DGP, 在改进了随机采样带来的偏置后, 在正负面情绪识别效果上均有进一步提高, F 值分别提高了 0.14% 和 0.37%。

实验表明, 若简单地将不同训练数据糅合在一起使用, 对情感分类性能反而可能有损失, 因此有必要实施迁移学习方法。对比现有的方法, 本文的方法超过了其他非迁移学习方法, 达到了最佳性能, 表明该方法通过引入辅助数据集, 有效提升了文本情感分析的效果。

5 结论

针对文本情感分析中存在的领域相关性和样本偏置问题, 本文提出在基于深度表示学习获得样本分布式表示基础上, 引入基于 *k* 近邻高斯过程迁移学习的思想, 以测试集为指导, 从辅助数据中提取额外数据和信息提高情感分类器的性能。在 COAE2014 微博情感倾向性识别数据集上的实验显示本文的方法可以有效降低领域相关性和样本偏置的影响, 有效提高情感分类性能。

参考文献:

- [1] Xu R.F, Wong K.F, Xia Y. Coarse-Fine Opinion Mining-WIA in NTCIR-7 MOAT Task[C]//Proceedings of NTCIR 2008: 307-313.
- [2] Kim Y. Convolutional Neural Networks for Sentence Classification[C]//Proceedings of EMNLP 2014:1746-1751.
- [3] Wang S, Manning C. D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification[C]//Proceedings of ACL 2012:90-94.
- [4] Taboada M., Brooke J., Tofiloski M., et al. Lexicon-based Methods for Sentiment Analysis [J]. Computational Linguistics, 2011, 37(2):267-307.
- [5] Fu G, Wang X. Chinese Sentence-level Sentiment Classification Based on Fuzzy Sets[C]//Proceedings of ACL 2010:312-319.
- [6] Bollegala D., Weir D., Carroll J. Cross-domain Sentiment Classification Using a Sentiment Sensitive Thesaurus[J]. IEEE TKDE, 2013, 25(8):1719-1731.
- [7] Pang B., Lee L., Vaithyanathan S. Thumbs up?: Sentiment Classification Using Machine Learning Techniques[C]//Proceedings of ACL 2002:79-86.

- [8] Zhao J, Liu K, Wang G. Adding Redundant Features for CRFs-based Sentence Sentiment Classification[C]//Proceedings of ACL 2008:117-126.
- [9] Socher R., Perelygin A., Wu J. Y., et al. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank[C]//Proceedings of EMNLP 2013:1631-1642.
- [10] Xiao S, Chengcheng L, Fuji R. Sentiment Analysis for Chinese Microblog based on Deep Neural Networks with Convolutional Extension Features[J]. Neurocomputing, 2016.
- [11] Pan S. J, Yang Q. A Survey on Transfer Learning[J]. IEEE TKDE, 2010, 22(10):1345-1359.
- [12] Dai W, Yang Q, Xue G. R, et al. Boosting for Transfer Learning[C]//Proceedings of ICML 2007:193-200.
- [13] Argyriou A., Evgeniou T., Pontil M. Convex Multi-task Feature Learning[C]//Proceedings of NIPS 2007:19-41
- [14] Zadrozny B. Learning and Evaluating Classifiers under Sample Selection Bias[C]//Proceedings of ICML 2004:114-121.
- [15] Fan W., Davidson I., Zadrozny B., et al. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias[C]//Proceedings of ICDM 2005:605-608
- [16] Blitzer J., McDonald R., Pereira F. Domain Adaptation with Structural Correspondence Learning[C]//Proceedings of ACL 2006:120-128.
- [17] Xu R.F, Xu J, Wang X. Instance Level Transfer Learning for Cross Lingual Opinion Analysis[C]//Proceedings of ACL 2011:182-188.
- [18] Gui L, Xu R.F, Lu Q, et al. Cross-lingual Opinion Analysis via Negative Transfer Detection[C]//Proceedings of ACL 2014:860-865.
- [19] Gui L, Lu Q, Xu R.F, et al. A Novel Class Noise Estimation Method and Application in Classification[C]//Proceedings of CIKM 2015:1081-1090.
- [20] Kandemir M. Asymmetric Transfer Learning with Deep Gaussian Processes[C]//Proceedings of ICML 2015:730-738.
- [21] Snelson E., Ghahramani Z. Sparse Gaussian Processes Using Pseudo-inputs[C]//Proceedings of NIPS 2005:1257-1264.



吴冬茵（1992-）硕士研究生，主要研究方向为自然语言处理，文本情绪计算。
Email:wudongyinhit@gmail.com



桂林（1988-）博士研究生，主要研究方向为机器学习，情感分析。
Email:guilin.nlp@gmail.com



陈钊（1990-）硕士，助理工程师，主要研究方向为复杂网络，信息传播，自然语言处理，文本情绪计算。
Email:gilbertchen@126.com



通讯作者

徐睿峰（1973-）博士，副教授，博士生导师，主要研究方向为自然语言处理，文本情绪计算，认知计算。
Email:xuruifeng@hitsz.edu.cn