基于情感轮和情感词典的文本情感分布标记增强方法

华 鑫　刘平生　曾雪强　左家莉　王明文
江西师范大学计算机信息工程学院

论文摘要

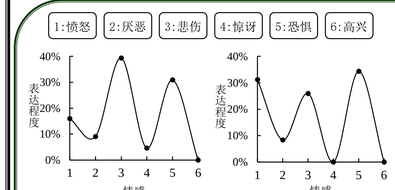
情感分布学习是一种近年提出的多情绪分析模型，通过为示例关联能同时记录多个情绪表达程度的情感分布，可以有效地处理具有情绪模糊性的情感分析任务。目前，情感分布学习面临的一个重要困难是缺乏已标注的文本情感分布数据集。为了利用已有的单标记情感数据集，情感分布标记增强方法可以用于将示例的情绪标签增强为情感分布。基于文本中的情感词蕴含着大量的情绪信息的特点，该文在引入普鲁契克情感轮心理学模型的基础上，提出基于情感轮和情感词典的情感分布标记增强方法（Emotion Wheel and Lexicon based emotion distribution Label Enhancement, EWLLE）。EWLLE方法基于情绪的心理学距离为句子的真实情绪标签和情感词的情绪标签分别生成离散高斯分布，然后通过分布的叠加将两种信息综合为统一的情感分布。在常用的四个文本情感数据集上的对比实验表明，在情绪识别任务上EWLLE方法的性能优于已有的情感分布标记增强方法。

相关工作

情感分布学习

情感分布学习（Emotion Distribution Learning, EDL）的建模目标是找到一个函数将句子$\boldsymbol{s}_i$映射为情感分布$\boldsymbol{d}_i=\{d_i^j\}_{j=1}^{C}$，其中$d_i^j$表示句子$\boldsymbol{s}_i$的第$j$种情绪的表达程度值，$d_i^j\in[0,1]$且$\Sigma_j d_i^j=1$。

如图1所示，SemEval数据集对句子的6种基本情绪的表达程度进行了标注。但大量已有的传统数据集只具有单标记情绪标签，具有情绪分布标记的数据集很少。



(a) Iraqi suicide attack kills two US troops as militants fight purge.　(b) Israelis retaliate after attack by Lebanese Army.

图1 SemEval数据集的两个例句及其在6种基本情绪上的表达程度

情感分布标记增强

在EDL中，情感分布标记增强的目标是将示例的情绪标签增强为情感分布，即将训练数据集中句子$\boldsymbol{s}_i$的情绪标签y_i扩展为情感分布$\boldsymbol{d}_i=\{d_i^j\}_{j=1}^{C}$。

基于情感词典的情感分布标记增强

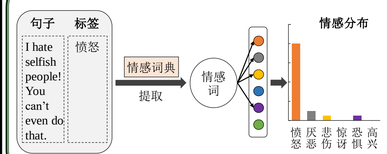


图2 基于情感词典的情感分布标记增强方法

Zhang等人于2018年提出了一种基于情感词典的情感分布标记增强方法（Lexicon based emotion distribution Label Enhancement, LLE）。如图2所示，例句的真实情绪标签是愤怒，其在LLE方法生成的情感分布中的得分最高。同时通过提取句子文本中的情感词的情绪标签，在情感分布中增加了厌恶、悲伤和恐惧三种得分较低的次要情绪。

LLE在句子的真实情绪标签之外，引入了情感词信息，但未考虑人类情绪的心理学先验知识，即没有利用情绪间存在的高度相关性。

基于Mikels情感轮的情感分布标记增强

Yang等人于2017年提出基于Mikels情感轮的情感分布标记增强方法（Mikels' emotion Wheel based emotion distribution Label Enhancement, MWLE）。MWLE方法基于Mikels情感轮计算情绪之间的距离，再采用高斯分布将情绪标签变换为情感分布。但是，由于没有利用句子文本中的情感词信息，MWLE方法的性能劣于LLE方法。到目前为止，还没有情绪分布标记增强工作同时考虑了心理学和语言学知识。

算法原理

本研究通过引入普鲁契克情感轮中的心理学知识和情感词包含的语言学信息，提出基于情感轮和情感词典的情感分布标记增强方法。EWLLE方法基于普鲁契克情感轮上的间隔角度定义情绪间的心理学距离，并基于此距离为真实情绪标签和情感词的情绪标签分别生成离散高斯分布，最后将两种标签的分布叠加为统一的情感分布。不同于已有的标记增强方法，EWLLE方法综合考虑了情绪的心理学和语言学知识，生成的情感分布包含更多的信息量。

普鲁契克情感轮

图3 普鲁契克情感轮

罗伯特·普鲁契克于1980年提出的情感轮理论，是一种从心理学角度描述人类情绪间相互关系的经典模型。普鲁契克情感轮包含4组对立的8种基本情绪，对角位置的情绪表现出对立性（负相关性），位置相邻的情绪具有相似性（正相关性）。

依据普鲁契克情感轮中相邻的情绪具有相似性的特点，我们定义两种情绪间的距离为其对应的在情感轮中的间隔角度。间隔角度越小的情绪距离越近，情绪每间隔45度距离定义为1。比如，高兴与信任是两个相邻情绪，间隔45度，对应的距离为1；高兴与悲伤间隔180度，对应的距离为4。

给定句子$\boldsymbol{s}_i$，EWLLE方法通过查找情感词典提取出文本中的情感词集合$\boldsymbol{w}_i=\{w_{i,k}\}_{k=1}^{n_i}$，其中$n_i$是句子$\boldsymbol{s}_i$的情感词数量。同时，每个情感词$w_{i,k}$有若干个关联的情绪标签$\{p_{i,k}^t\}_{t=1}^{m_k}$，其中$m_k$是$w_{i,k}$的情绪标签数量。一个句子的文本中可以包含多个情感词，也可能没有情感词；每个情感词至少关联一个情绪标签。

给定情绪标签α，$\alpha\in\{1,2,\cdots,C\}$，我们认为应该按照以下两个准则为其生成情感分布：1）情绪标签α对应的得分应该是情感分布中的最大值，以确保真实情绪在分布中的主导地位；2）其他情绪的得分应随着离α的距离的增大而减小，以使得和真实情绪越相似的情绪在分布中的得分越高。因为基本情绪在情感轮上是环状关系，生成的情感分布应该是一个以真实情绪标签α为中心并左右对称递减的分布。

基于上述两个准则，EWLLE方法假设从情绪标签生成的情感分布服从正态分布，并采用离散高斯分布将情绪标签α扩展为分布$\boldsymbol{f}_\alpha=\{f_\alpha^a\}_{a=1}^{C}$。具体而言，以如下公式计算真实情绪标签$\alpha$为中心的离散高斯分布$\boldsymbol{f}_\alpha$：

$$f_\alpha^a=\frac{1}{\sigma\sqrt{2\pi}Z}\exp\left(-\frac{|a-\alpha|^2}{2\sigma^2}\right),\qquad(1)$$

式中，σ是离散高斯分布的标准差，Z是归一化因子，使得$\sum_a f_\alpha^a=1$，$|a-\alpha|$是情绪a与真实情绪α之间的情感轮距离，采用基于普鲁契克情感轮的方式计算。标准差σ越大，情绪离散分布越平坦，考虑的情绪范围越大；反之，考虑的情绪范围越小。本文实验中离散高斯分布的标准差σ设为1。

基于公式1，EWLLE方法为句子$\boldsymbol{s}_i$的真实情绪标签y_i生成高斯分布$\boldsymbol{f}_{y_i}$，并为每个情感词的情绪标签$p_{i,k}^t$生成高斯分布$\boldsymbol{f}_{p_{i,k}^t}$。然后，EWLLE方法将代表两种信息的高斯分布$\boldsymbol{f}_{y_i}$和$\boldsymbol{f}_{p_{i,k}^t}$叠加起来，得到综合的情感分布$\boldsymbol{d}_i$。

$$\boldsymbol{d}_i=\frac{1-\lambda}{\sum_{k=1}^{n_i}m_k}\cdot\sum_{k=1}^{n_i}\sum_{t=1}^{m_k}\boldsymbol{f}_{p_{i,k}^t}+\lambda\cdot\boldsymbol{f}_{y_i},\quad(2)$$

式中，n_i是句子$\boldsymbol{s}_i$的情感词数量，m_k是句子$\boldsymbol{s}_i$的第k个情感词$w_{i,k}$的情绪标签数量，$\boldsymbol{f}_{p_{i,k}^t}$是情感词$w_{i,k}$的第$t$个情绪标签$p_{i,k}^t$的高斯分布，$\boldsymbol{f}_{y_i}$是真实情绪标签$y_i$的高斯分布，真实情绪标签$y_i$的权重系数$\lambda$用于控制分布$\boldsymbol{f}_{y_i}$在情感分布$\boldsymbol{d}_i$中的比例。

实验结果

本文采用4个在文本情绪识别中常用的单标记英文数据集（TEC、Fairy Tales、CBET和ISEAR）作为实验数据集。情感分布学习情绪识别模型，采用被证明具有较好性能的多任务卷积神经网络（Convolutional Neural Network, CNN）模型。标记增强方法生成的情感分布作为训练样本的监督信息使用。CNN模型预测的情感分布中得分最高的情绪，作为情绪识别的预测结果。

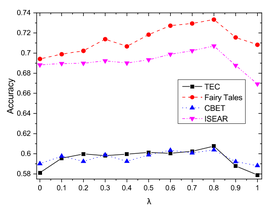


图4 真实情绪标签权重系数λ变化情况下EWLLE方法在4个数据集上的情绪识别准确率

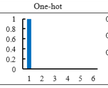
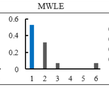
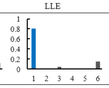
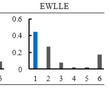
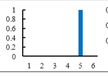
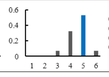
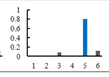
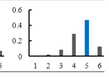
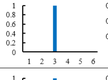
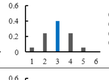
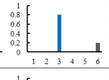
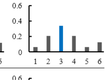
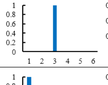
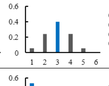
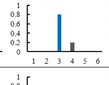
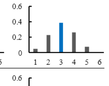
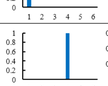
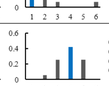
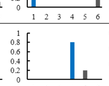
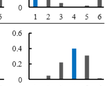
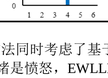
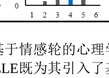
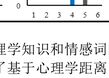
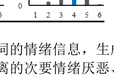
表1 4种布标记增强方法的多任务CNN情绪识别的情感分精确率、召回率、F1分数和准确率

数据集	方法	Pre(%)	Rec(%)	F1(%)	Acc(%)
TEC	One-hot	52.32	41.96	43.77	56.74
	MWLE	55.19	42.03	44.16	57.89
	LLE	55.56	42.89	45.85	59.23
	EWLLE	**57.33**	**43.08**	**47.77**	**60.76**
Fairy Tales	One-hot	72.53	62.01	65.19	69.16
	MWLE	73.25	63.51	65.54	70.83
	LLE	75,45	**64.27**	66.33	71.51
	EWLLE	**77.98**	63.56	**67.72**	**73.33**
CBET	One-hot	61.33	56.77	57.98	56.74
	MWLE	62.56	58.71	58.19	58.80
	LLE	63.17	58.69	58.28	59,71
	EWLLE	**63.72**	**59.44**	**59.06**	**60.38**
ISEAR	One-hot	66.78	65.21	64.49	65.46
	MWLE	68.32	66.59	65.84	66.93
	LLE	70.21	67.86	67.78	68.14
	EWLLE	**72.89**	**69.03**	**68.17**	**70.71**

从图4可以看出，所有4个数据集表现出了相似的准确率变化趋势。虽然不同数据集上的情绪识别准确率绝对得分有较大的差异，但都是当λ=0.8时准确率达到最优。λ取值在0到0.7之间时，情绪识别的准确率逐步上升；说明此时增加真实情绪标签的权重是有益的。当λ=0.8时，情感词和真实情绪标签的信息量达到平衡，情绪识别准确率获得最优值。当λ>0.8时，准确率随着λ增大而快速下降；表明句子文本中的情感词信息对情绪识别非常重要，当情感词信息不足时标记增强方法的性能会显著降低。

从表1的结果可以看出，本文提出的EWLLE方法在所有4个数据集上取得了比其它标记增强方法总体更好的结果。以ISEAR数据集的准确率为例，EWLLE方法的准确率比LLE方法高出2.57%，比MWLE方法高出3.78%，比One-hot方法高5.25%。相对于对比模型中性能最好的LLE方法，除了在Fairy Tales数据集上LLE的召回率优于EWLLE外，EWLLE方法在所有4个数据集上的4个性能指标上均有明显的提升。这一实验结果说明，在使用真实情绪标签和情感词信息生成情感分布之外，增加考虑情感的心理学先验知识有益于提高情绪识别性能。

表2 4种情感分布标记分布方法生成的情感分布对比（句子文本中的情感词用粗体标出，情感分布中的真实情绪标签用蓝色标出，情绪标签1：愤怒、2：厌恶、3：悲伤、4：惊讶、5：恐惧、6：高兴）

序号	句子文本及其标签	情感词及其标签	生成的情感分布 One-hot	MWLE	LLE	EWLLE
1	I **just absolutely lost** my **mind**. 标签: 1	just: 6 absolutely: 6 lost: 3 mind: 6				
2	**People** are more likely to **die** in **hospital** at **the weekend**. 标签: 5	people: 6 die: 3, 5 hospital: 3, 5 the: 6 weekend: 6				
3	I'm **exhausted but** can't sleep. I **just want** my **nap**. 标签: 3	exhausted: 3 but: 6 just: 6 want: 6 nap: 6				
4	Packing to **leave** tomorrow. 标签: 3	leave: 3, 4				
5	4 of 6 speedways in Midland don't **have** pizza 标签: 1	have: 6				
6	Arabic coursework **deadline** extended by a week. 标签: 4	deadline: 5				

如表2所示，本文提出的EWLLE方法同时考虑了基于情感轮的心理学知识和情感词的情绪信息，生成的情感分布更为合理。比如，例句1的真实情绪是愤怒，EWLLE既为其引入了基于心理学距离的次要情绪厌恶、悲伤和高兴，同时基于情感词信息调高了次要情绪高兴的权重。与LLE的做法类似，EWLLE方法对基于情感词的情绪信息赋予了较低的权重。总体而言，EWLLE生成的情感分布具有比MWLE和LLE更好的合理性，本文4.3节的情绪识别实验结果也说明EWLLE方法具有更优的性能。

论文结论

- 本文提出了一种基于情感轮和情感词典的情感分布标记增强方法（Emotion Wheel and Lexicon based emotion distribution Label Enhancement, EWLLE），用于将单标记文本情感数据集中的情绪标签增强为情感分布。
- EWLLE方法基于普鲁契克情感轮心理学模型为句子的真实情绪标签和情感词的情绪标签分别生成离散高斯分布，然后将两种高斯分布叠加为统一的情感分布。
- 不同于已有的标记增强模型，EWLLE方法综合考虑了情感心理学知识和情感词的语言学信息。
- 实验结果表明，EWLLE方法在情绪识别任务上的性能优于已有的情感分布标记增强方法。

在下一步的工作中，我们将考虑在情感分布标记增强方法中引入其他的情感先验知识，并尝试多种不同的情感建模方式以更有效地利用先验知识。

西安电子科技大学 XIDIAN UNIVERSITY