

基于深度学习的网络心理咨询问题智能多分类模型

程会林 曾 伟 杨琳琳 岑萧萍 蒋琳琳

通讯作者：魏航

(广州中医药大学医学信息工程学院, 广东 广州 510006)

摘 要: 为了简化网络心理咨询分诊流程和降低咨询助理的劳动强度, 提出一种基于深度学习的网络心理咨询问题智能多分类模型。采用Word2vec对来访者的自述情况进行文本特征提取, 设计卷积神经网络的深度学习分类器, 实现对来访者咨询问题的智能分类。与传统机器学习分类方法相比, 所提方法的准确率提升约20%。本研究为网络心理咨询问题的智能分诊提供了一种新方法。

关键词: 网络心理咨询; 自述分类, Word2vec; 卷积神经网络

1 引言

随着互联网技术的不断发展, 网络心理咨询逐渐发展起来, 仅“525心理网”便已解决近857万个心理问题。心理咨询网站一般配备有一名咨询助理, 根据来访者的自述情况进行分类, 匹配到合适的心理咨询师。然而, 这种分诊的准确性与效率依赖于咨询助理的经验知识。为减轻咨询助理的负担, 本文提出一种对咨询问题进行智能分诊的方法。传统的分类方法需要进行复杂的处理来进行特征提取, 再根据提取的特征来训练分类器。卷积神经网络通过结构重组和减少权值将特征提取和分类功能进行融合, 这种结构比以往多种算法性能更为高效^[7]。因此, 本文拟利用文本挖掘技术, 设计卷积神经网络多分类模型, 对来访者的自述情况进行智能分类, 简化分诊流程。

2 实验步骤与分析

2.1 流程图

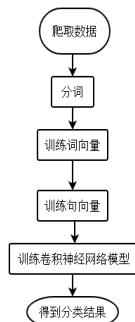


图1 流程图

2.2 数据预处理

为获得数据, 训练分类模型以供使用, 本文利用八爪鱼采集器从“咨询中国网”中爬取“公益心理问答模块”数据, 共得到6个类别, 分别为“婚恋情感”“亲子关系”“人际关系”“认识自我”“压力疏导”及“子女成长”。但每一类数据的数量不同, 为使训练时类别均衡, 本文将“亲子关系”“人际关系”及“子女成长”合并为“关系成长”类。最终共有4个类别, 其中婚恋情感自述情况条数310条, 关系成长自述情况条数284条, 认识自我自述情况条数310条, 压力疏导自述情况条数310条。

利用结巴分词(jieba)将爬取的数据分词, 因文本并不长, 所以未加载停用词, 例如“我觉得我活得很累”被分为了“我觉得我活得很累”。

将分词后结果转为txt文本, 代入Word2vec模型^[1], 得到256维的词向量, 即一个中文词(t)由256个数字表示, 表示为R(t)。为训练出更好的词汇模型, 本文使用约43万条, 涉及影视、小说、心理、司法内容的文本进行训练, 得到新的Word2vec结果。

$R(t) = \text{Word2vec}(t)$

接着利用训练得到的Word2vec模型, 计算每条自述情况的向量 $R(d_j)$, 即句子 d_j 中n个词汇的Word2vec向量的均值, 表示为:

$$R(d_j) = \frac{\sum \text{Word2vec}(t)}{n} \quad \text{where } t \in d_j$$

2.3 模型构建

以256维的文本向量作为特征, 分类标签作为label, 分别使用决策树^[2]、随机森林^[3]、SVM^[4]及卷积神经网络^[5]进行建模。

将一条文本所得的256维向量看作 16×16 的二维矩阵。将 $32 \times 5 \times 5$ 的小型矩阵, 作为第一层的卷积核, 对于大矩阵中的每个值, 计算其周围的值与卷积核对应位置的乘积, 将结果相加最终得到的终值则为新矩阵该位置的值。

一层卷积的操作可概括为:

(1) 文本向量通过多个不同的卷积核的滤波, 并加偏置, 提取出局部特征, 每一个卷积核映射出一个新的2D文本。

(2) 将前面卷积核的滤波输出结果, 进行非线性的激活函数ReLU处理。

(3) 对激活函数的结果进行最大池化, 保留最显著的特征, 实现数据压缩, 并提升模型的畸变容忍能力。

在第一层卷积层中, 共有 $32 \times 5 \times 5$ 的卷积核, 第二层为 $64 \times 5 \times 5$ 的卷积核, 利用relu作为激活函数, 全连接得到自述情况的分类。并利用AdamOptimizer不断优化得到结果, 经过20000次迭代后, 得到准确率最高的模型。

2.4 实验结果

表1 准确率结果

模型	训练准确率	测试准确率
SVM	0.9988	0.5178
决策树	1.0	0.5863
随机森林	0.9170	0.0106
卷积神经网络	0.8269	0.7918

从实验结果可以看出, 相较于传统分类方法, 卷积神经网络在短文本的分类中更具优势。准确率得到大大提升。卷积神经网络通过卷积核的权值共享, 减少了参数个数, 亦提取了显著特征, 通过梯度下降算法的参数更迭, 使得模型的准确率得以大大提升。

2.5 实验进阶

在得到关于自述情况的分类后, 可得知其咨询的问题所属大类为何, 即可推荐擅长此大类的心理咨询师。而实际情况中, 推荐时若能够更加细致, 则心理咨询师的推荐将能够更加个性化。根据自述情况中来访者的期望, 自动筛选出重要信息, 例如希望心理咨询师性别为男, 年龄为30岁以上等等。

本文简单化处理, 对于心理咨询师的要求, 统一在文本的最后, 表述为: “心理咨询师要……”, 便可使用正则表达式, 获取其对咨询师的要求, 接着由“性别……”“……岁”等模式的正则, 继续细化出要求得到的结果, 由机器直接获取信息, 按照心理咨询师的标签, 进行推荐。

3 结语

与传统方法对比, 卷积神经网络在短文本分类上训练效果更好, 但准确率仍有提升的空间。若能考虑到关键词在分类过程中的显著作用, 相信能够得到更好的结果。例如通过“老公”, “男朋友”两词, 即可判定此文本属于婚恋情感类。但可以肯定的是, 卷积神经网络在心理咨询自述情况上的分类是有应用价值的。当配合以正则表达式, 便可更个性化推荐心理咨询师。

参考文献

- [1] 王奕森, 夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术, 2018, 12(01): 49-55.
- [2] 郭东亮, 刘小明, 郑秋生. 基于卷积神经网络的互联网短文本分类方法[J]. 计算机与现代化, 2017(04): 78-81.
- [3] 唐明, 朱磊, 邹显春. 基于Word2Vec的一种文档向量表示[J]. 计算机科学, 2016, 43(06): 214-217+269.
- [4] 张棣, 曹健. 面向大数据分析的决策树算法[J]. 计算机科学, 2016, 43(S1): 374-379+383.
- [5] 崔建明, 刘建明, 廖周宇. 基于SVM算法的文本分类技术研究[J]. 计算机仿真, 2013, 30(02): 299-302+368.
- [6] 许闯. 网络心理咨询五步法[J]. 政工导刊, 2017(07): 40-41.
- [7] 夏从零, 钱涛, 姬东鸿. 基于事件卷积特征的新闻文本分类[J]. 计算机应用研究, 2017, 34(04): 991-994.