doi: 10.3969/j.issn.0490-6756.2019.01.010

基因遗传算法在文本情感分类中的应用

邓昌明, 李 晨, 邓可君, 张治坤, 袁 玲, 姜 宁,
彭一明, 邢承杰, 卞 晶, 陈 光, 王梦淑, 王雪琴
(北京大学计算中心, 北京 100871)

摘 要: 本文以微博文本为主要实验对象, 提出适合卷积神经网络进行自我优化的编码方式,分别将每一层看做是一个染色体,将每一层中的参数看做是一个基因片段,采用混合双重非数值编码的方式编码每个 CNN 框架,设计出适合于 CNN 网络的选择、交叉和变异的算法,并且把基因遗传算法(GA)和与卷积神经网络相结合,提出了基于情感分析算法的遗传算法(GA-CNN). 通过对传统算法与 GA-CNN 的实验与对比分析,良好地展示了自我优化性.

关键词: 基因算法; 情感分析; 深度学习; 自我进化
中图分类号: TP391.1; TP183 **文献标识码**: A **文章编号**: 0490-6756(2019)01-0045-05

Application of genetic algorithm in text sentiment classification

DENG Chang-Ming, LI Chen, DENG Ke-Jun, ZHANG Zhi-Kun,
YUAN Ling, JIANG Ning, PENG Yi-Ming, XING Cheng-Jie,
BIAN Jing, CHEN Guang, WANG Meng-Shu, WANG Xue-Qin
(Computer Center, Peking University, Beijing 100871, China)

Abstract: In this paper, we use Sina Weibo text as the main experimental dataset, and propose a coding method suitable for self-optimization of convolutional neural networks. Our coding method encodes each CNN framework using a hybrid double non-numeric encoding by treating each layer as a chromosome and the parameters in each layer as a gene segment respectively, and selection, crossover and mutation algorithms are devised for CNN networks. We also propose a genetic algorithm based on sentiment analysis algorithm (GA-CNN) which combines genetic algorithm (GA) with convolutional neural network. The experiment and comparative analysis of GA-CNN and traditional algorithms demonstrates the self-optimization of our method.

Keywords: Genetic algorithm; Sentiment classification; Deep learning; Self-optimization

1 引 言

随着网络技术的进步和社会应用的普及,网页的交互信息越来越多的被企业、政府所重视. 基于网页的信息获取、挖掘、分析也被逐渐提升到了国家安全的高度. 网页信息的交互包含浏览历史记录、跳转路径、发布的信息、微博、视频、语音以及注册的个人信息、账号等等,他们包含每个人的部分或者全部核心信息,如个人的工作、情感、生活、经济、习惯和信仰等等. 对网页数据的挖掘与分析将有助于个人乃至国家的发展. 本文主要以网页数据中的微博为主要例题,对其中所表露出来的情感进

收稿日期: 2018-06-25
作者简介: 邓昌明(1985-), 河南周口人, 硕士, 工程师, 研究方向为网络与数据库技术. E-mail: dengcm@pku.edu.cn
通讯作者: 张治坤. E-mail: zhangzhikun@pku.edu.cn

行分析研究,并对算法自我优化的可行性进行分析探讨.

情感分析(Sentiment Analysis,SA)又称为倾向性分析和意见挖掘,它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程,其中情感分析还可以细分为情感极性(倾向)分析,情感程度分析和主客观分析等[1]. 情感极性分析的目的是对自然语言中多包涵的正向情绪、负向情绪和中立情绪进行判别. 大多数应用场景中,只分为两类. 例如对于"喜欢"和"讨厌"这两个词,表达的就是两种相反的情感. 情感分析在建立完善互联网的舆情监控系统,对异常或突发事情的检测以及心理学、社会学、金融预测等领域中都有广泛应用.

目前国内外对于微博等短文本的情感挖掘分析已经做出了很多研究[2]. 常用的方法如朴素贝叶斯[3],逻辑回归[4]、K 最近邻分类 KNN 算法(k-NearestNeighbor)[5]、支持向量机(Support Vector Machine, SVM)[6,7] 和卷积神经网络(Convolutional Neural Network,CNN)[8]等,都在不同的目标对象的情况有良好的表现. 但是对于不同的任务和不同的数据源,如中文微博和英文微博,对文字微博和表情微博等的分析仍有较大差异[9,10]. 针对不同的任务,人们会人工尝试不同的算法并通过调整优化来实现最佳匹配和提升效率. 对于参数结构众多,探索空间巨大的情况,这种方式不仅时间效率较低,而且探索空间局限,优化效果不明显. 因此能够让算法自我进化,并且在全局空间内进行自我优化,不仅能够节省人力,还能够提升算法对不同任务的适应性,在现实工作中具有较强的现实意义[11].

本文主要以中文微博数据为例,以情感分析为主要实验对象,结合遗传算法(Genetic Algorithm, GA),实现对情感分析算法的自我优化,提出了以卷积神经网络为对象的遗传进化算法(GA-CNN),并通过实验,来模拟实现对中文情感分析算法的自我进化过程和结果.

2 传统方法情感分析实验

微博以不超过 140 字为一个表达方式,具备词语种类丰富、语句简短、主题发散及创新词语多等特点,相对于长文本而言,在情感分析的问题上面临的问题和困难更多[12]. 文本情感分析过程一般包括文本预处理、情感特征提取和情感分类等步骤. 文本预处理指对文本进行分词,对词性进行标注,以及停用词的成立等操作;情感特征的提取是指按照一定的规则,把具有明显倾向性的单元要素从微博文本进行抽取的过程;情感分类是利用抽取出来的情感特征对文本进行区分,对主观性文本极性和强度进行分类. 中文微博情感分类大致上包括:基于情感词典的分类方法和基于机器学习的分类方法两类[9].

2.1 实验环境

本文中的所有实验均在如表 1 所示的实验环境中完成.

表 1 实验环境及其配置

Tab. 1 Experimental environment and configuration

实验环境	环境配置
操作系统	linux Ubentu 14.04
CPU	i7
内存	16G
显卡	GeForce GTX 1080
GPU	8G
编程语言	Pethon 2.7
分词工具	ICTCLAS
深度学习框架	Keras 1.1.2+TensorFlow 0.12
传统开发框架	skiLearn 0.18.1

2.2 数据集的选择与处理

试验数据来源于新浪微博的数据集. 该数据集包含 1.6 万余语句,其中 1.2 万来自于 PC 端,0.4 万条来自移动端. 将来自 PC 端的数据进行分类,按照心理学对情感的归类,将"happiness"、"like"归为正向情感("pos");将"anger"、"disgust"、"fear"归为负向情感("neg");将"surprise"、"none"归为中性情感("none"). 并通过约 20 人进行独立认证,采用最高的归类,进行划分.

同时将来自移动端的数据被标记直接标注为"正向情感"、"负向情感"和"无情感"3 个类别. 数据的标记过程仍然采用原先汇总人员进行独立标注,标注中忽略了表情符号所表达的情感,仅对中文自然语言所表达出的情感进行了标记,选取其中比例最高的标注作为单条语句的情感类型. 情感类型分为三类,正向情感、负向情感和中性情感,其中正向情感语句 4699 条,负向情感语句 4891 条,中性情感语句 6548 条. 采用 80%进行训练,20%进行测验.

2.3 实验结果对比分析

上述传统方法和基础 CNN 方法在实验环境中的测试结果如表 2 所示.

表 2　传统算法与 CNN 算法的对比

Tab. 2　Comparison between traditional algorithm and CNN algorithm

模型	训练精确度	测试精确度	分类	训练集	测试集
朴素贝叶斯	0.5297	0.2199	3	12910	3228
逻辑回归	0.5714	0.2331	3	12910	3228
SVM-线性	0.4705	0.2385	3	12910	3228
SVM-非线性	0.9951	0.5354	3	12910	3228
SVM-RBF -GridSearch	0.9948	0.5355	3	12910	3228
CNN-static	0.9952	0.4642	3	12910	3228
CNN-rand	0.9865	0.5268	3	12910	3228
CNN-non-static	0.9958	0.535	3	12910	3228

从试验可知，对于传统分类算法而言，SVM 的性能较高，在该数据情况下，朴素贝叶斯算法的精确度较低，其次是逻辑回归以及线性 SVM 算法.

对于深度学习的 CNN 网络，在这个样本集中，表现出了较好的分类效果. 该试验中的 CNN 分别进行了三类试验，分别是基于预训练词向量的 CNN-static、随机编码的 CNN-rand 和经过调参的 CNN-non-static. 并分别对 CNN 进行了人为调整参数. 结果显示 CNN-non-static 比最好的 CNN-rand 高出 0.009，达到了 53.5%. 但相对于传统的情感分析分类算法，CNN-none-static 比 SVM-RBF-GridSearch，精确度效果却并没有提升，甚至还低 0.05%.

进一步分析说明对于该 CNN 的网络结构和参数的设定并没有达到 CNN 网络的最大性能，同样对于 SVM 的算法也并没有达到其最大的精确度. 那么对于 CNN 这样网络结构复杂，层次可以无限加深，探索空间巨大的情况，人为调参仅能实现局部性搜索优化，无法实现最优或近似最优的效果优化. 而对于网格搜索而言，它是一种枚举型收索，它的特点是耗时长，全局性差. 对于深度学习的自我探索，谷歌在 2017 年进行了研究，Barret Zoph[13] 等人于 2017 年初尝试了一种基于大型服务阵列上的自我遍历探索优化的尝试，实验结果完成了基于 RNN 的图像识别的自我增强优化. 但这样的自我优化需要较大的资源，对于普通算法的或者资源有限的前提下，需要寻找一个有效的算法进行高效的全局性的自动调整优化. 结合 Barret Zoph 等人的探索，本文提出了基于卷积神经网络的遗传进化算法(GA-CNN).

3　基于基因遗传算法的自我优化算法

本文中，采用 CNN 探索模型进行基于遗传算法的优化，主要讨论该模型是否能够通过模拟进化完成结构性和参数性的探索，以达到根据不同任务和数据源进行自我结构和参数的变更，使性能达到最优.

3.1　基因遗传算法相关理论

遗传算法 GA 是 1975 年由美国 Michigan 大学的 Holland 教授在其专著《自然界和人工系统的适用性》中首先提出的. 遗传算法，也称进化算法，是受达尔文的进化论的启发，借鉴生物进化过程而提出的一种启发式搜索算法. 借鉴生物进化论，遗传算法将要解决的问题模拟成一个生物进化的过程，通过复制、交叉、突变等操作产生下一代的解，并逐步淘汰掉适应度函数值低的解，增加适应度函数值高的解. 这样进化 N 代后就很有可能会进化出适应度函数值很高的个体[14,15].

3.2　GA-CNN 算法的设计

CNN 网络结构中，可以讨论的参数和结构很多. 在 GA-CNN 的算法探索中，将每一层网络结构看作是一个染色体. GA-CNN 算法的系统架构如图 1 所示；其整体流程如算法 1.

算法 1 GA-CNN 算法

Begin

步骤 1　对数据进行规范处理并分为训练集、评价集和测试集；

步骤 2　初始化 CNN 框架结构种群，预先设定最大迭代次数 G，当前种群代数 $g=1$；

步骤 3　对 CNN 种群中的每个框架结构进行学习训练；

步骤 4　用评价集对训练的 CNN 模型，进行评估，获得 CNN 框架结构种群所对应的适应度；

步骤 5　采用轮盘赌法生成交配目标；

步骤 6　对交配目标进行交叉操作，并进行训练评估适应度；

步骤 7　利用变异操作，对交叉结果进行变异，并进行训练评估适应度；

步骤 8　判断新产生的结果是否优于交配目标，更新 CNN 结构种群，更新对应的适应度；

步骤 9　如果 $g<G$ 且不满足收敛条件，$g=g+1$，转到步骤 5，否则转到步骤 10；

步骤 10　输出精英个体模型作为最终的分类模型.

End.

其中,GA−CNN 算法中的交叉变异算法流程如图 2 所示;其交叉变异逻辑流程如算法 2.

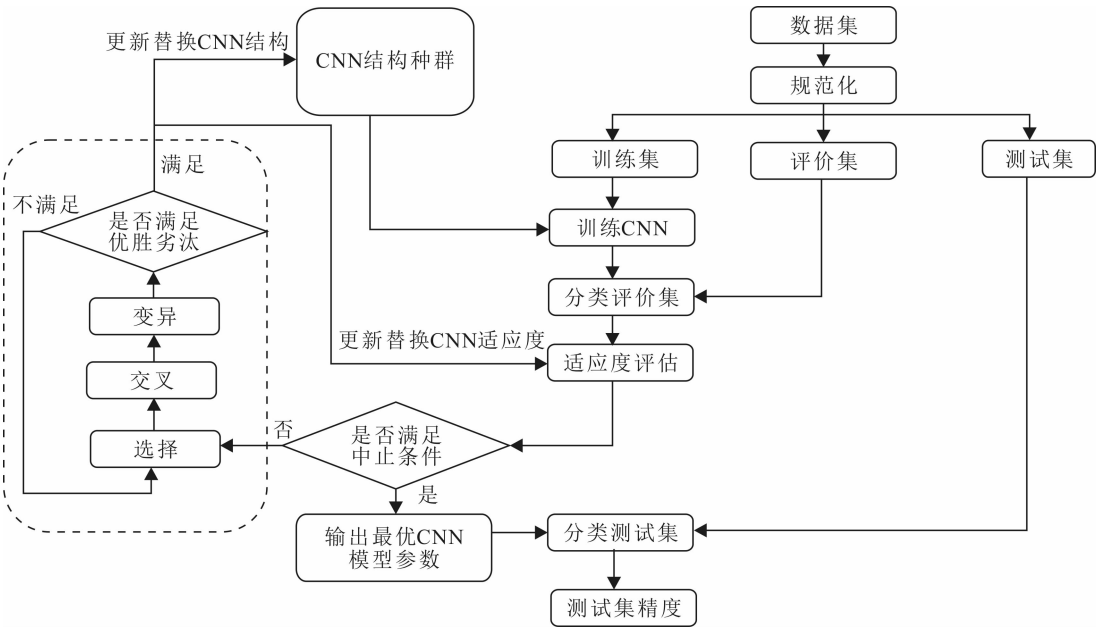

图 1 GA-CNN 算法的系统架构

Fig. 1 The architecture of GA-CNN algorithm

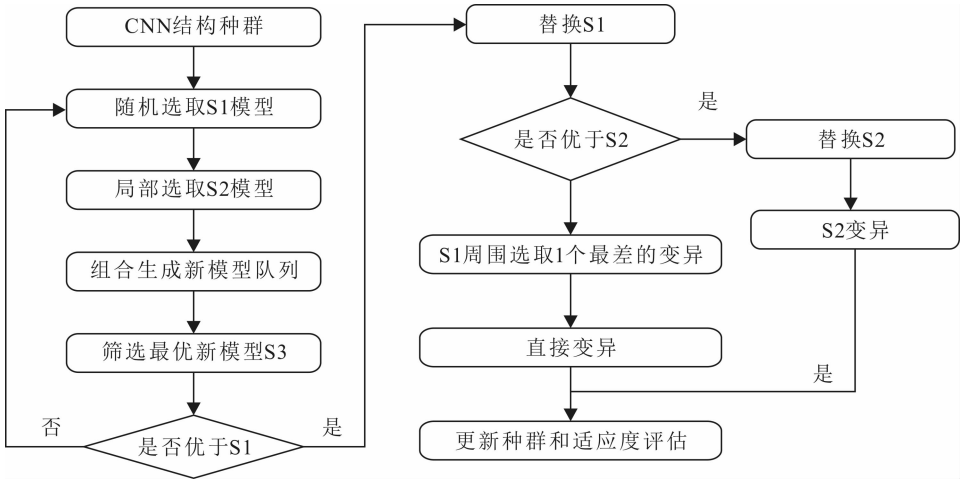

图 2 GA-CNN 算法中的交叉变异算法逻辑

Fig. 2 The logic of cross mutation in GA-CNN algorithm

算法 2 GA-CNN 算法交叉变异逻辑

Begin

步骤 1 采用随机法在 CNN 种群中选取基模型 S1;

步骤 2 在 S1 周围局部选取,距离为 1 的交配模型 S2;

步骤 3 交叉产生新的模型队列,对产生的新模型进行训练学习,评估其适应度;

步骤 4 比较筛选适应度最高的模型 S3;

步骤 5 判断新产生的模型 S3 是否优于基模型 S1,如果优于 S1,替换 S1;如果不优于 S1,舍弃,转到步骤 1;

步骤 6 判断是否优于交配模型 S2;优于交配模型 S2,转到步骤 7;不优于交配模型 S2,转到步骤 8;

步骤 7 替换 S2,接着 S2 变异,转到步骤 9;

步骤 8 在 S1 周围选取一个适应度最差的进行变异;

步骤 9　更新种群和适应度评估.

End.

GA-CNN 算法与传统 CNN 测试后的结果对比如表 3 所示.

表 3　GA-CNN 算法与 CNN 算法的对比

Tab. 3　Comparison between GA-CNN algorithm and CNN algorithm

模型	训练命中度	测试命中度	分类	训练集	评估集	测试集
CNN-static	0.9952	0.2385	3	12910		3228
CNN-rand	0.9865	0.5354	3	12910		3228
CNN-non-static	0.9958	0.5355	3	12910		3228
GA-CNN	0.9916	0.7708	3	10328	2582	3228

综上所述，可以看出 GA-CNN 算法，经过进化，进行有效的自我调优，调整了自己的结构和模型参数，提升了模型准确性，从 52.68% 上升到了 77.08%. 该进化在 85 次时达到了收敛，取得了一个近似最优解.

4　结　论

实验分析，GA-CNN 算法有效地解决了人为调参数的局限性，对分布空间广，探索空间大的 CNN 模型架构以及参数能够有效的探索和自动优化，在探索时间和空间上都相对人为调参有较大提升. 相对于枚举法而言具有较好的收敛性.

但该算法也存在一定的问题和思考：由于资源空间有限，对基因和染色体种类的模拟具有局限性，大量参数和变数引入可能带来较大的影响. 同时对于染色体的编码由于种类较少，类似于二进制编码. 初始化的种群结构不同，可能带来的进化时间成本和结构都有所不同. 最后的结果可能在最大迭代次数 G 完成时，仍只能得到一个近似最优解，而这个近似最优解可能存在差异.

参考文献：

[1]　蒋延华. 风景油画创作的情感分析 [J]. 美术教育研究，2012，2012：25.

[2]　王文华，朱艳辉，徐叶强，等. 基于 SVM 的产品评论属性特征的情感倾向分析 [J]. 湖南工业大学学报，2012，26：76.

[3]　陈红玉. 数据挖掘中贝叶斯分类算法的研究 [J]. 光盘技术，2009，2009：57.

[4]　周志华. 机器学习 [M]. 北京：清华大学出版社，2016.

[5]　贾可亮，樊孝忠，许进忠. 基于 KNN 的汉语问句分类 [J]. 微电子学与计算机，2008，2008：156.

[6]　马波. 支持向量机多类分类算法的分析与设计 [D]. 扬州：扬州大学，2008.

[7]　饶刚. 支持向量机(SVM)算法的进一步研究 [D]. 重庆：重庆大学，2012.

[8]　张建明，詹智财，成科扬，等. 深度学习的研究与发展 [J]. 江苏大学学报：自然科学版，2015，36：191.

[9]　任小燕. 中文情感分析综述 [J]. 科技信息，2011，31：202.

[10]　周胜臣，瞿文婷，石英子，等. 中文微博情感分析研究综述 [J]. 计算机应用与软件，2013，30：161.

[11]　Wei W，Gulla J A. Sentiment learning on product reviews via sentiment ontology tree [C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.[s. l.]：ACM，2010.

[12]　王岩. 基于共现链的微博情感分析技术的研究与实现 [D]. 北京：国防科学技术大学，2011.

[13]　Zoph B，Le Q V. Neural architecture search with reinforcement learning [EB/OL]. (2017-02-15). [2018-05-28]. https://wenku. baidu. com/view/3080c3392379168884868762caaedd3383c4b52b. html.

[14]　王晓天，边思宇. 基于遗传算法和神经网络的 PID 参数自整定 [J]. 吉林大学学报：理学版，2018，56：953.

[15]　陈龙. 基于遗传算法的约束性多 TSP 问题及其应用 [J]. 重庆邮电学院学报：自然科学版，2000：67.

引用本文格式：

中　文：邓昌明，李晨，邓可君，等. 基因遗传算法在文本情感分类中的应用 [J]. 四川大学学报：自然科学版，2019，56：45.

英　文：Deng C M，Li C，Deng K J，*et al*. Application of genetic algorithm in text sentiment classification [J]. J Sichuan Univ：Nat Sci Ed，2019，56：45.