基于Adaboost框架下自动编码器提升方法的文本分类

山东科技大学电气与自动化工程学院　刘广秀　宋单单

【摘要】针对文本分类问题，使用深度学习算法中的自动编码器模型网络建造Adaboost框架下的弱分类器，并且在自动编码器神经网络部分引入噪声，引入神经元间歇性工作机制，更改不同参数以及层数构造弱分类器。利用神经网络的稀疏性提高分类器的泛化性，加入Adaboost框架实现深度学习的集成算法。

【关键词】SAE；Adaboost；文本分类；激励函数

0 引言

大数据时代的到来，网络的普及，信息量呈爆炸性趋势增长，人们迫切需要一种实用性技术来有效的地组织和管理信息。从大量的信息中获取有效信息变得尤为重要。文本挖掘、自然语言处理、信息检索等技术很好地解决了信息过载时代的文本数据管理问题，文本分类技术作为这些领域的重要基础，在近年来得到了快速发展和广泛关注[1]。文本分类的方法有很多，典型的有朴素贝叶斯分类器[9]、BP神经网络分类器、K近邻算法（KNN）、支持向量机（SVM）分类器等，这些分类器在文本分类中均取得了很好的效果。并且在传统分类器的使用上，有很多学者提出了改进方案，使得分类效果有所提升。比如基于深度信念网络的文本分类器算法[10]，基于稀疏编码器的文本分类算法[7]等。深度学习作为一种新兴的多层神经网络降维算法，通过组建含有多个隐层的神经网络深层模型，对输入的高维数据逐层提取特征，以发现数据的低维嵌套结构，形成更加抽象有效的高层表示[8]。传统BP神经网络梯度越来越稀疏，易于收敛于局部最优，有标签的训练数据类别涵盖不全，且类别比例差别较大，使用深度学习网络很易产生过拟合问题。根据已有的深度学与boosting结合案例，本文提出Adaboost与编码器深度学习算法相结合算法。结合深度学习网络提取特征良好的特点，本文提出使用深度学习网络中的SAE网络作为Adaboost框架下的弱分类器，使用不同激励函数等参数变换构造不同的自动编码器网络，加入Adaboost框架的思想实现深度学习集成算法。

1 Adaboost算法

Boosting算法是一种把若干个分类器整合为一个分类器的方法，能够将预测精度仅比随机猜度略高的弱学习器增强为预测精度高的强学习器,这在直接构造强学习器非常困难的情况下,为学习算法的设计提供了一种有效的新思路和新方法.作为一种元算法框架,Boosting几乎可以应用于所有目前流行的机器学习算法以进一步加强原算法的预测精度,因此应用十分广泛,产生了极大的影响。Boosting方法有许多不同的变形，更具一般性的AdaBoost形式由ROBERT E.SCHAPIRE和YORAM SINGER在1999年提出，其核心思想是针对同一个训练集训练不同的分类器，然后把这些弱分类器集合起来，构成一个更强的最终分类器[5]。Adaboost的算法流程如下：

第1步：给定一组具有标签的训练数据集：

$T=\{(x_1, y_1),\cdots,(x_N, y_N)\}$

其中：$x_i \in X, y_i \in Y=\{1,-1\}$。

第2步：初始化训练数据的权值分布。每一个训练样本最开始时都被赋予相同的权值：1/N。

$$D_1=(w_{11},w_{12},\cdots w_{1i},\cdots,w_{1N}),\qquad w_{1i}=\frac{1}{N},\ i=1,2,\cdots,N$$

第3步：使用弱学习算法，迭代m=1,2,3...M次。

a. 使用具有权值分布的全训练集，进行基本元分类器C_m训练得到h_m，或按照权重w_{1i}对训练集进行采样后对元分类器C_m训练得到分类器h_m。

b. 计算C_m在训练数据集上的分类误差率公式：

$$e_m=P(h_m(x_i)\neq y_i)=\sum_{i=1}^{N} w_{mi}I(h_m(x_i)\neq y_i)$$

即C_m在训练集上的误差率就是被C_m分类错误的样本的权值之和。

c. 计算弱分类器C_m的权值系数公式：

$$\alpha_m=\frac{1}{2}\ln\frac{1-e_m}{e_m}$$

d. 更新训练数据集的权值公式：

$$D_{m+1}=(w_{m+1,1},w_{m+1,2}\cdots w_{m+1,i},\cdots,w_{m+1,N})$$

$$w_{m+1,i}=\frac{w_{mi}}{z_m}\exp(-\alpha_m y_i h_m(x_i)),\ i=1,2,\cdots,N$$

Z_m是规范化因子，使得D_{m+1}为一个概率分布：

$$Z_m=\sum_{i=1}^{N} w_{mi}\exp(-\alpha_m y_i h_m(x_i))$$

第4步：组合各弱分类器得到最终分类器表达式：

$$H(x)=sign[\sum_{m=1}^{M}\alpha_m h_m(x)]$$

上述式子组成了Adaboost算法的基本步骤。Adaboost算法的自适应性在于：前一个基本分类器分错的样本会得到加强，加权后的全体样本再次被用来训练下一个基本分类器。

2 降噪稀疏自动编码器

基本自动编码器的描述如下：自动编码器是运用了反向传播进行无监督学习的神经网络，学习的目的就是输出信号尽可能复现输入信号。为了实现这种复现，自动编码器就必须捕捉可以代表输入数据的最重要的特征，就像主成分分析那样，找到可以代表原信息的主要成分[4]。基本的自动编码器接收输入向量$\mathbf{x}$，在激活函数的作用下对其进行线性变化，得到一个编码结果$\mathbf{y}$[3]。本文选取sigmoid函数作为激活函数，计算公式如下：

$\mathbf{y}=f_\theta(\mathbf{x})=s(W\mathbf{x}+\mathbf{b})$，$\mathbf{z}=g_{\theta'}(\mathbf{y})=s(W'\mathbf{y}+\mathbf{b}')$。

$\theta=\{W,\mathbf{b}\}$为编码参数，$\theta'=\{W',\mathbf{b}'\}$为解码参数。其中$W$是一个$d'\times d$的权重矩阵。$W'$为$W$的转置矩阵，$\mathbf{b}$和$\mathbf{b}'$是偏置向量。

稀疏自动编码器是加上一些约束条件得到的新的Deep Learning方法。在AutoEncoder的基础上加上L1的Regularity限制（L1主要是约束每一层中的节点中大部分都要为0，只有少数不为0），我们就可以得到Sparse AutoEncoder法[2]。

SAE损失函数表达式：$L(x,W)=\|Wy-x\|^2+\lambda\sum_i|y_i|$

降噪自动编码器是在自动编码器的基础上，在训练数据中加入噪声，所以自动编码器必须学习去除这种噪声而获得真正的没有被噪声污染过的输入[7]。因此，这就迫使编码器去学习输入信号更加鲁棒性的表达，这也就促使了它的泛化能力比一般编码器强。DA可以通过梯度下降算法去训练。

3 基于Adaboost算法和降噪稀疏自动编码器的文本分类模型

本文以DSAE（降噪稀疏自动编码器）为弱分类器基本原型[8]，调整层数以及激励函数种类构造不同条件下的弱分类器，使用NLPIR分词系统提取文本特征，使用TFIDF作为词语的权值，根据该权值来选择特征词，并统计词频作为文本特征训练集。整个算法的流程图如图1所示。

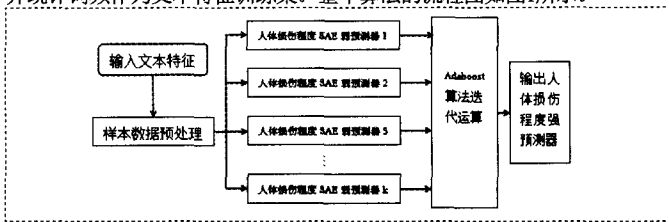

图1 基于Adaboost算法和SAE网络的人体损伤程度预测流程

（下转第197页）

```
flag=0;
}
}
}
```

5.2 数据滤波程序

数据滤波程序用于将获取的数据进行稳定运算，使输出的数据更加平滑，减少误差。由于传感器对外每秒钟输出十次数据，而对屏幕数据每秒钟刷新一次，所以采用定时器定时处理获取到的数据。

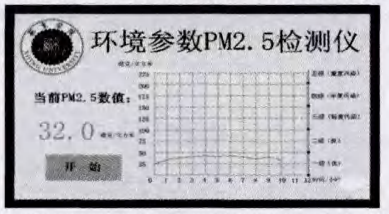

图4 环境参数PM2.5检测仪界面

6 HMI界面设计

HMI界面设计使用VisualTFT上位机软件。VisualTFT上位机软件为DCOLOUR（广州大彩）公司配套HMI显示器开发使用，直接使用内嵌的虚拟串口屏与单片机进行通讯，也支持Keil环境下直接调试，与真实硬件操作一样。图4为环境参数PM2.5检测仪在西安某地区十小时监测数据。

7 总结

本文利用高性能嵌入式系统和灰尘传感器以及HMI串口显示器，设计了简易的PM2.5检测仪。该检测仪运用了夏普灰尘传感器，其参考值出厂已标定，串口数据输出更加稳定可靠，在数据处理中采用嵌入式系统，能够及时进行数据的处理、减数据误差，显示采用HMI串口显示器，通过对显示界面及通讯的单独设计，减少设计难度提高系统稳定性。实验证明，该检测仪能够较好的满足普通环境下的环境参数PM2.5的获取，达到设计要求。

参考文献

[1]陈权昌,李兴富.单片机原理及应用[M].广州:华南理工大学出版社,2007:84-102.

[2]李庆亮.C语言程序设计实用教程[M].北京:机械工业出版社,2005:32-58.

[3]康华光.电子技术基础数字部分[M].北京:高等教育出版社,2008:203-209.

[4]杨欣.电子设计从零开始[M].北京:清华大学出版社,2005:28-102.

作者简介：

姚冲（1994—），陕西西安人，大学本科，现就读于西安西京学院。

侯新刚（1982—），陕西西安人，讲师，主要研究方向为电路设计。

李红波（1982—），陕西西安人，讲师，主要研究方向为数据采集与控制。

（上接第194页）

注册成功成为网站会员后，可以查看新闻动态，查看魅力声音并去进行留言，还可以修改自己在网站注册的资料信息。需要与后台数据新闻表（xw）、留言表（ly）、用户姓名（name）表进行连接绑定。

(12)网站投票子页面设计。网站投票栏目用户可以对网站整体印象进行投票，并查看投票结果。需要与后台数据库投票表（vote）连接绑定。

6.2 后台设计

网站后台只有系统管理员可以进入。后台的设计页面比较简单，易于操作。后台页面都是动态页面，都使用ASP+ Dreamweaver以及JavaScript语句。

(1)后台登录页面设计。需要输入用户名、密码以及验证码，正确后方可登录到后台管理页面。需要与管理员数据表（admin）进行连接绑定。

(2)后台新闻管理页面设计。新闻管理可以对新闻进行浏览、修改、删除以及发表。需要与新闻数据表（xw）进行连接绑定。

(3)后台留言管理页面设计。留言管理可以对留言进行审核、回复以及删除。需要与留言数据表（ly）数据进行连接绑定。

(4)后台用户管理页面设计。用户管理可以对普通用户进行资料的修改和删除以及对管理员的资料修改。需要与用户姓名表（zc）和管理员数据表（admin）数据进行连接绑定。

(5)后台图片管理页面设计。图片管理可以添加前台需要展示的图片信息。需要与图片数据表（tupian）进行连接绑定。

7.创新点

网站前后台的页面设计，前台网站栏目很有特色，都是围绕“魅力一中”主体展开，网站可实现注册、登录、新闻的上传、发表留言、上传图片以及对本网站进行投票打分等功能。本网站建设有静态部分也有动态部分，充分体现数字技术的开放性、交互性和共享性的特征。

通过完整制作一个网站，我感到要想做成一件事情就要有严谨认真的态度，还要有不断完善、精益求精的精神。

（上接第195页）

设计基于SAE网络弱预测器[6]：每个分类器可能出现的不同特征设计：加入稀疏惩罚项、不加入稀疏惩罚项、神经网络预训练的激发函数使用sigmoid函数、神经网络训练数据是否加入噪声、神经网络监督微调部分使用激发函数sigmoid函数、神经网络监督微调部分使用tanh函数、神经网络的层数变化、节点数变化。以变化自动编码器参数等方式，实现每个分类器的结构互异性，加大各个分类器的分类各异性，实现网络的结构设计使得不同弱预测器具有不同的预测倾向性，运用Adaboost集成各个弱分类器加大集成分类器的泛化性，使得分类器分类效率更高。

参考文献

[1]M.S.Bartlett,G.Littlewort,M.G.Frank,C.Lainscsek,I.Fasel,and J.R.Movellan.Recognizing facial expression:Machine learning and application to spontaneous behavior.In CVPR,volume 2,pages.568-573,2005.

[2]J.Mairal,F.Bach and J.Ponce.Sparse Modeling for Image and Vision Processing.Foundations and Trends in Computer Graphics and Vision.vol 8.number 2-3.pages 85-283.2014.

[3]孙志军,薛磊,许阳明.深度学习研究综述[J].计算机应用研究,2012,29(8):2806-2810.

[4]Hinton,G.E.and Salakhutdinov,R.R.Reducing the dimensionality of data with neural networks.Science 2006.

[5]曹莹,苗启广,刘家辰,高琳.AdaBoost算法研究进展与展望[J].自动化学报,39(6):745-758.

[6]Bengio,Y.,Lamblin,P.,Popovici,P.,Larochelle,H.Greedy Layer-Wise Training of Deep Networks.NIPS 2006.

[7]Pascal Vincent,Hugo Larochelle,Yoshua Bengio and Pierre-Antoine Manzagol.Extracting and Composing Robust Features with Denoising Autoencoders.ICML 2008.

[8]殷力昂.一种在深度结构中学习原型的分类方法[D].上海:上海交通大学,2012.

[9]张雪峰.设计贝叶斯分类器文本分类系统[J].电脑知识与技术,2005(20).

[10]陈翠平.基于深度信念网络的文本分类算法[J].计算机系统应用,2015,24(2):121-126.

阅读此文的还阅读了：

1. 基于降噪自动编码器的中文新闻文本分类方法研究
2. 文本分类中基于AdaBoost.MR的改进中心法
3. 基于AdaBoost和Krawtchouk矩的刀具形状分类识别
4. 基于概率潜在语义分析和Adaboost算法的文本分类技术研究
5. 神经模糊技术在车型自动分类中的应用
6. 孙子兵法主题研究述论
7. 基于Rocchio方法和k均值聚类的支持向量机文本分类方法
8. 一种基于文本分类的知识树自动构建方法
9. 基于CCIPCA，LSSVM的文本自动分类算法
10. 基于词语上下文关系的文本自动分类方法研究
11. 实验室废弃物料收集装置的改进
12. 挖掘课内外资源 提升学生的文化素养
13. 基于词语上下文关系的文本自动分类 方法研究
14. 基于CBR的文本自动分类研究
15. 一种基于词上下文向量的文本自动分类方法
16. 基于规则的自动分类在文本分类中的应用

17. 基于模糊向量和BP网络的Web文本自动分类方法

18. 基于AdaBoost的文本隐写分析

19. 基于粗糙集和RBF神经网络的文本自动分类方法

20. 基于改进VSM的Web文本分类方法

21. 辨性质 明角度 趋大流——谈谈影视时代小说文本的分类

22. 用Excel实现自动编班

23. 基于SVM的网络文本信息自动分类

24. 对中国《石油天然气资源/储量分类》标准的评论与建议

25. 基于加权模糊推理网络的文本自动分类方法

26. 基于知识树的文本自动分类方法探索

27. 基于AdaBoost特征约减的入侵检测分类方法

28. 基于序列的文本自动分类算法?

29. 基于SVM的网络文本信息自动分类

30. 文本类名中的中国古代结构观念探析——以《梦溪笔谈》、《营造法式》为例

31. 基于综合特征的Bp-adaboost工业仪表图像分类方法

32. 模糊聚类分析在文本分类中的应用

33. 基于PCA-AdaBoost的舌象颜色分类研究

34. 一种基于粗糙-神经网络的文本自动分类方法

35. 在对话中提升学生的人文素养:以《送行》为例

36. 基于Adaboost的电子邮件分类算法

37. 支付账户分类监管:经验借鉴与政策框架

38. 基于Web的新闻文本分类技术的研究

39. 基于Adaboost-BP神经网络的图像情感分类方法研究

40. 广义粗糙参数与性能关系系统分类研究

41. 基于SVM的中文文本自动分类研究

42. 中文文本的关键词自动抽取和模糊分类

43. 一种基于Adaboost.M1的车型分类算法

44. 基于栏目的藏文网页文本自动分类方法

45. 基于EXCEL内置函数的记录表的新设计方法

46. 基于文本内容的农业网页信息抽取和分类研究

47. 基于AdaBoost的文本隐写分析

48. 基于AdaBoost组合学习方法的岩爆分类预测研究

49. 基于KNN的中文文本自动分类研究

50. 旋转刀塔刀位编码器的研究