面向网络安全的有害信息智能识别算法研究

庞雪茹
（西安航空职业技术学院 陕西 西安 710089）

摘要：针对网络上存在危害社会稳定的有害信息难以被准确监控分析的问题，为提高网络安全的预防能力，文中进行了网络有害信息智能识别算法的研究。首先，利用人工智能中的KNN和SMOTE算法进行有害信息的数据获取、扩充，为后续模型训练提供必要的样本数据；然后，通过信息增益进行特征提取，并使用词袋模型进行格式转化；最后，利用堆叠降噪自编码器模型学习特征向量中隐含的信息，进而实现有害信息的智能识别。通过多次对比测试实验结果表明，文中提出的有害信息智能识别算法，具有较高的识别率，平均识别率为83.12%，证明了该方法的有效性。

关键词：有害信息识别；网络安全；人工智能；堆叠降噪自编码器
中图分类号：TN99；TP183 **文献标识码：**A **文章编号：**1674-6236(2020)18-0071-05
DOI：10.14022/j.issn1674-6236.2020.18.016

Design of intelligent identification algorithm of harmful information for network security

PANG Xue-ru
(*Xi'an Vocational and Technical College of Aeronautics and Astronautics*, *Xi'an* 710089, *China*)

Abstract: In order to improve the prevention ability of network security, this paper studies the intelligent recognition algorithm of network harmful information. Firstly, KNN and smote algorithm in artificial intelligence are used to acquire and expand the data of harmful information, providing necessary sample data for the subsequent model training; secondly, feature extraction is carried out through information gain, and format transformation is carried out by word bag model; finally, the hidden information in the feature vector is learned by stack noise reduction self encoder model to realize harmful information Intelligent identification. The experimental results show that the proposed algorithm has a high recognition rate, the average recognition rate is 83.12%, which proves the effectiveness of the scheme.

Key words: identification of harmful information; internet security; artificial intelligence; stacked noise reduction autoencoder

随着移动互联网的迅速发展，其影响力逐步赶超传统互联网。智能手机等多种智能终端是移动互联网的窗口，使得越来越多的人生活在更紧密的空间，时刻享受着网络大数据带来的服务[1-3]。然而，在海量的网络信息中，有时会夹杂着敏感信息，甚至带有恐怖主义性质的内容。若得不到及时的处理，这些不良信息会通过移动互联网大范围传播，影响社会稳定[4-7]。例如，在新西兰克赖斯特彻奇所发生的枪击案中，袭击过程被枪手以网络直播的形式在全世界快速传播。引起了大范围的民众恐慌情绪，助长了恐怖主义的传播，造成严重的国际影响。因此，应当加强网络尤其是移动互联网中敏感信息的识别和监管[8-11]。

针对网络空间的有害、敏感信息识别，国内外科研学者取得了重要的成果[12-16]。美国TDT项目实现了网络数据流中重要信息的归纳整理，为敏感信息的识别提供了技术积累；Autonomy针对中国互联网

收稿日期：2019-12-11 **稿件编号：**201912096
基金项目：2019年陕西高校辅导员工作研究(2019FKT35)
作者简介：庞雪茹(1995—)，女，陕西西安人，硕士，助教。研究方向：反恐法学。

环境，推出了政府网络信息检测产品；同时，北大方正技术研究院在网络舆情预警方面，进行了舆情预警辅助决策、支持系统的研究等。

本文针对网络中涉及恐怖主义、血腥内容的有害信息，进行主题分类，通过两次分类识别的方式提升识别率。利用人工智能中的KNN和SMOTE算法，进行有害信息的数据获取、扩充。通过信息增益，进行文本信息的特征提取，并使用词袋模型将文本格式转化成模型可识别的格式。最后，利用机械学习中的堆叠降噪自编码器模型学习特征向量中隐含的信息，进而实现有害信息的智能识别。

1 有害信息的识别任务特点和总体框架

由于互联网中的信息有文字、图片、音频、视频等多种形式，均有可能混杂着有害、敏感信息。为了尽可能拦截全部有害信息，需要对不同形式的网络内容进行监管和识别。首先，需要对有害信息进行定义。根据网络上的内容，有害信息涉及以下方面，如图1所示。

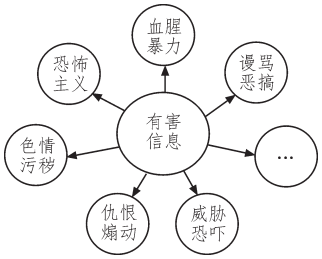

图1 有害信息的种类

根据上文对有害信息的分析，本文针对有害信息的智能识别算法，采用两次识别方式：1）实现主题分类，主要分为政治、娱乐、社会、教育、经济和其他；2）对各主题的具体有害信息进行识别。由于网络信息数量巨大、形式多样，人工智能技术被用来进行数据挖掘。通过提取到的样本数据，利用机器学习进行相应的算法与模型的构建，以此实现有害信息的智能识别。

2 基于人工智能的有害信息识别算法

2.1 有害信息智能识别算法结构

人工智能是一种企图了解人类学习方式并实现模仿、延伸的技术，其核心是机器学习。通过模拟人类的学习行为来获得新的知识和能力，并对已有的知识架构进行更新。根据应用需求，利用人工智能所实现的功能过程，主要涉及4个要素：数据、特征、模型和算法。其之间的关系，如图2所示。由应用需求确定数据的类型与范围，并利用人工智能技术进行数据的获取；根据机器学习编写适用于应用的算法，进行数据的特征提取和模型构建；将特征导入至构建好的模型中，训练以学习数据中蕴含的特征规律，进而实现特定的功能。数据与特征是人工智能实现功能的基石，其决定了功能的上限；而所构建的模型与算法，是机器学习实现功能的途径，模型与算法的优劣决定着到达上限的快慢。

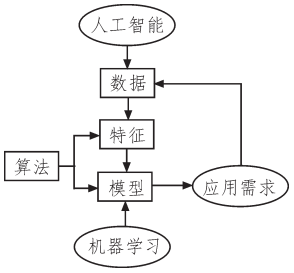

图2 智能识别算法结构示意图

2.2 数据的确定和扩充

K最近邻分类算法（KNN算法）被用来实现5个主题的分类。KNN算法是一种基本分类与回归方法，常被用于人工智能应用中。距离度量、K值的选择和分类决策规，则是KNN算法的基本要素。通过收集网络上各个社交软件、媒体软件的信息，制作成训练数据集T：

$$T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\} \tag{1}$$

上式中，x_i为采集数据的特征向量，是训练的输入；y_i为数据的类别，是训练的输出。具体过程如下：

1）根据事先设定的距离度量公式，在数据集中寻找与预选值x最邻近的K个点，记为$Mk(x)$；

2）在$Mk(x)$中，依据设定的分类决策规则来判定x的分类y，y可表达为：

$$y=\arg\max_{c_j}\sum_{x_i\in M_k(x)}I(y_i=c_j) \tag{2}$$

当y_i=c_i时，I=1，否则I=0。在KNN算法中，数据的相似程度使用距离度量表示。本文采用欧式距离作为距离的计算方式，具体如下：

$$L_2(x_i,x_j)=\left(\sum_{l=1}^{n}\left|x_i^l-x_j^l\right|^{0.5}\right) \tag{3}$$

对于本文中几种有害信息主题的样本x，分别从k个最邻近集合中选择n个样本，记作x_i。i的取值范围为[1，n]，n的数值需要根据样本数据集的不平衡

率进行选取。在x与x_i之间通过随机线性插值的方式增加样本数量，具体方式如式(4)：

$$x_{new}=x+rand(0,1)\times(x_i-x) \tag{4}$$

式(4)中，$rand(0,1)$表示的是在0~1之间的任意一个随机数。值得注意的是，随机数的使用，使得插入的数值不一定属于少数类样本。同时，也未排除可能出现的噪声影响。针对上述问题，对SMOTE算法做出一定的改进。假定$Cand(X)$为最后能与样本X随机线性插值的样本集合；$Noise(X)$为样本集中的噪声。首先，计算X与训练集合中其他样本的欧式距离，从中选取最近的k个样本，构成集合$A(X)$；再从中选取k个欧式距离最近的少数类样本，构成集合$B(X)$。利用下式，计算样本X的k值邻近集合中，少数类所占的比例：

$$r=\frac{\left|A(X)\bigcap B(X)\right|}{\left|A(X)\right|} \tag{5}$$

式(5)中，使用$|A(X)|$来表示集合$A(X)$中元素的数量，分子表示集合$A(X)$与集合$B(X)$的交集元素个数。本文中有害信息的主题分类为6，则k的取值为6。比例r的取值可分为以下几种情况：

R=0时，令$Cand(X)$为空集，$Noise(X)$为X；

R=0.2时，令$Cand(X)$为空集，$Noise(X)$为空集；

R=0.3或0.5时，则有式(6)：

$$Cand(X)=A(X)\bigcap B(X) \tag{6}$$

$$Noise(X)=B(X)-Cand(X) \tag{7}$$

当$Cand(X)$为非空集时，可从$Cand(X)$中随机挑选出n个样本与样本X，进行随机线性插值；而当$Noise(X)$为非空集时，需要从样本集合中删掉$Noise(X)$中的元素。因此，可以避免噪声的干扰下，增加少数类数据样本。

2.3　基于深度学习的有害信息的智能识别

上文进行了数据的扩充以便增加有害信息的比例，使得其数据特征与有害信息的类别比配更加充分。深度学习是机器学习的一种，通过模拟人脑的学习行为，以构建深层次模型的方式挖掘数据背后的特征。本文使用深度学习中的层叠降噪自动编码器神经网络，来实现数据特征的提取。

自动编码器是由高维度的输入数据、输出数据及低纬度的编码矢量组成，如图3所示。其中，编码矢量也被称为隐藏层。首先，对输入数据编码处理，经过激活函数后，再进行解码。在这一编码、解码过程中，由节点权重所构成的矩阵互为转置矩阵，使得输出信息与输入信息保持一致。降噪自动编码器的工作原理，是将具有一定特征的噪声引入样本数据中。经过编码、解码，统计输出结果中未受噪声影响的样本。因此，便可筛选出具有抗拒噪声的样本数据。

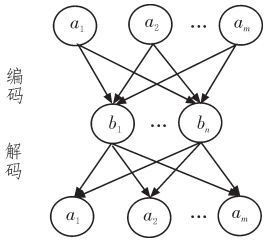

图3　自动编码器结构示意图

文本数据特征提取是通过信息增益进行的。信息增益在信息熵的基础上，来表征该特征词划分的纯度大小。其中，样本集X中第k类样本所占比例为P_k，则该样本集X的信息熵的表达式如式(8)：

$$Ent(X)=-\sum_{k=1}^{|y|}p_k\log_2(p_k) \tag{8}$$

式中，$|y|$表示需要判定的类别数量，$Ent(D)$的数值越大，表明该样本集的纯度越低。

样本集X信息增益的表达式如式(9)：

$$IG(X,t)=Ent(X)-\sum_{n=1}^{N}\frac{\left|X^n\right|}{\left|X\right|}Ent(X^n) \tag{9}$$

其中，T表示当前使用的特征词，N为t的分类个数。信息增益数值越大，表明通过该特征词划分，来实现纯度提升的概率越大。文本特征提取完后，通过词袋模型将文本格式转换为可数值计算格式，以便模型的迭代训练。

反向传播算法常被用于单层隐藏层的自动编码器中。然而，此类算法并不适用于具有多层隐藏层的模型，尤其在训练数据十分有限的情况下。上文虽然对有害信息通过插值增加了样本数量，但数量仍处于相对较低的比例。针对此类情况，本文将每一层隐藏层看作最基本的降噪自编码器，以无监督预训练的方式，逐层堆叠；再通过整体的反向微调优化，来实现有害信息的精准识别。具体结构，如图4所示。

3　测试与验证

为了验证本文所提出方案的有效性，针对不同主题的有害信息，进行算法测试与验证。本文测试样本数据采用中文与英文两种语言，分别选自人民

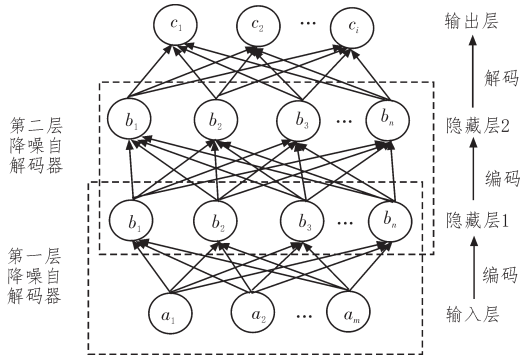

图4　含有两层隐藏层的堆叠降噪自编码器示意图

日报语料库和路透社语料库。每种语言，均按照5个主题各选择50段文档，具体如表1所示。

表1　6类主题样本数据

序号	文档类别	文档数量	有害信息文档数量	占比
1	政治	50	7	14%
2	娱乐	50	9	18%
3	社会	50	7	14%
4	教育	50	8	16%
5	经济	50	10	20%
6	其他	50	6	12%

从表1可以看出，各个主题分类的有害信息所占比例并不一致。存在比例过低，会导致识别精度较差。经改进SMOTE算法插值后有害信息样本数量，如表2所示。

表2　经改进SMOTE算法插值后6类主题样本数据

序号	文档类别	文档数量	有害信息文档数量	占比
1	政治	50	10	20%
2	娱乐	50	13	26%
3	社会	50	11	22%
4	教育	50	10	20%
5	经济	50	13	26%
6	其他	50	9	18%

有害信息识别测试实验，使用了Eclipse开发平台，运行环境为Linux。利用上文的样本数据，通过信息增益进行特征提取、词袋模型进行格式转换，生成特征词向量。分别进行两次测试实验：1）经改进后的SMOTE算法插值样本，与未经过插值的样本识别结果进行对比；2）基于堆叠降噪自编码器的深度学习模型，与浅层学习模型的有害信息识别进行对比。图5展示了不同特征向量维度、降噪自编码器层数有害信息识别率的对比。样本特征向量的维度，在一定程度上对识别率有着显著的作用，从10维度增长到40维度时，识别率增长较快；而超过40维度时，增长缓慢。降噪自编码器的层数增多，也会使有害信息识别率增大。

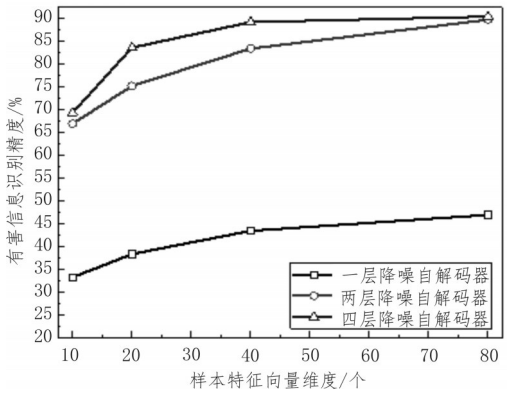

图5　不同特征向量维度、不同降噪自编码器层数有害信息识别率对比

图6展示了有无SMOTE算法插值对有害信息识别率的影响。改进SMOTE算法的使用，使得有害信息的平均识别率为83.12%，明显高于未进行过插值的样本。这表明，改进SMOTE算法可有效增加样本中少数类，即有害信息的数量。图7展示了堆叠降噪自编码器与浅层学习模型的有害信息识别率对比。从图中可以看出，堆叠降噪自编码器的有害信息识别率明显高于浅层学习模型。

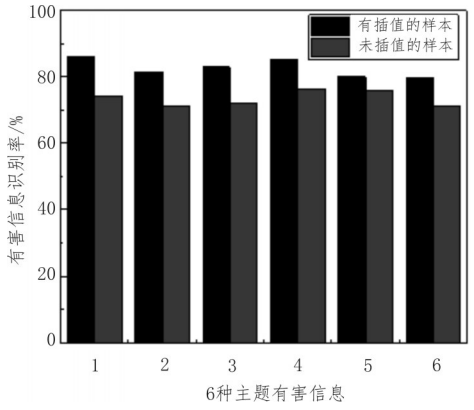

图6　6种有害信息主题类别的识别率对比

4　结　论

本文针对网络中涉及恐怖主义、血腥内容的有害信息。利用人工智能中的KNN和SMOTE算法，进行有害信息的数据获取、扩充。通过信息增益进行文本信息的特征提取，以及使用词袋模型进行格式转化。最后，利用机械学习中的堆叠降噪自编码器

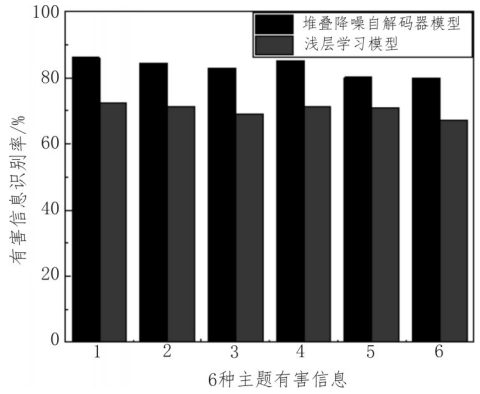

图7　堆叠降噪自编码器与浅层学习模型的有害信息识别率对比

模型学习特征向量中隐含的信息，进而实现有害信息的智能识别。通过多次对比测试实验表明，本文提出的有害信息智能识别算法，具有较高的识别率，证明了文中方案的有效性，可用于网络安全的防护。

参考文献：

[1] 吴良.基于大数据平台的信令数据采集技术研究[J].电子科技，2019，32(5):89-91，95.

[2] 罗煜.基于移动互联网技术的配调移动服务平台设计[J].电子设计工程，2018，26(23):48-51，56.

[3] 李磊.数据通信网络安全维护策略探讨[J].软件，2018，39(7):191-193.

[4] 龚文全.人工智能在有害信息识别服务的应用和发展趋势[J].电信网技术，2018(2):10-14.

[5] 陈传银.基于云计算平台的复杂网络分析关键算法研究[D].成都:电子科技大学，2014.

[6] 王莉，郑婷一，李明.网络媒体大数据中的异构网络对齐关键技术和应用研究[J].太原理工大学学报，2017，48(3):453-457.

[7] 文孟飞，刘伟荣，胡超.网络媒体大数据流异构多模态目标识别策略[J].计算机研究与发展，2017，54(1):71-79.

[8] 吴珊，李跃新.智能设备网络虚假信息行为识别与控制技术研究[J].计算机测量与控制，2019，27(4):88-91，133.

[9] 卢刚.面向网络社区的敏感信息语义计算方法研究[D].北京:北京邮电大学，2018.

[10]闫光辉，张萌，罗浩，等.融合高阶信息的社交网络重要节点识别算法[J].通信学报，2019，10(3):109-110.

[11]耿磊.无线网络中具有网络监管功能的端到端安全通信方案[D].西安:西安电子科技大学，2017.

[12]范亮，陈倩.人工智能在网络安全领域的最新发展[J].中国信息安全，2017(12):104-107.

[13]刘延华，高晓玲，朱敏琛，等.基于数据特征学习的网络安全数据分类方法研究[J].信息网络安全，2019，10(4):50-56.

[14]李雄飞，李军，董元方，等.一种新的不平衡数据学习算法PCBoost[J].计算机学报，2012(13):2202-2207.

[15]王一大.基于堆叠降噪自编码器的情感分类及其并行化研究[J].电信技术，2017，12(8):8-12.

[16]梁政.面向在线社交网络舆情的信息传播分析关键技术研究[D].长沙:国防科学技术大学，2014.

(上接第70页)

构建与教学实践[J].实验科学与技术，2016(6):209-211.

[11]任武，吴运新，许志杰，等.一种改进最小二乘复频域方法及其应用[J].华中科技大学学报(自然科学版)，2014(5):30-33.

[12]张新贺，金明录.空间调制信号的改进M-ML检测算法[J].大连理工大学学报，2016(2):140-146.

[13]李立，冯贺，赵建周.基于40 Gbit/s 4QAM-OFDM的RoF相干传输系统特性分析[J].郑州大学学报(理学版)，2017(2):119.

[14]李擎，全厚德，陈明.基于软件无线电的教学实验平台设计与实现[J].实验室研究与探索，2010，29(12):41-44.

[15]邢鑫，赵慧.基于LabVIEW和USRP的软件无线电通信实验平台设计[J].实验技术与管理，2016，236(5):160-164.

[16]戴伏生.软件无线电的通信系统实验平台研制(1)—件资源[J].实验室研究与探索，2019，38(8):66-70.