本文引用格式：曾光辉.基于匹配自主学习的网络信息识别与分类算法[J].自动化与信息工程,2024,45(3):45-50.
ZENG Guanghui. Network information recognition and classification algorithm based on matching autonomous learning[J]. Automation & Information Engineering, 2024,45(3):45-50.

基于匹配自主学习的网络信息识别与分类算法*

曾光辉
（广州工程技术职业学院，广东 广州 510900）

摘要：为提升网络信息的识别与分类准确率，针对海量网络信息的高维、高噪等特点，提出基于匹配自主学习的网络信息识别与分类算法。首先，利用支持向量机对网络信息进行识别；然后，通过奇异值分解算法构建检索矩阵进行奇异值分解、相关性查询；接着，计算网络信息的相似性匹配度，并将匹配度较高的网络信息输入到卷积神经网络中进行学习、训练；最后，输出网络信息分类结果。实验结果显示，该算法网络信息识别准确率达到 97.90%以上，针对不同类别网络信息的平均分类准确率为 98.04%，证明了该算法在实际应用中的有效性。

关键词：匹配自主学习；网络信息；支持向量机；奇异值分解；卷积神经网络；识别与分类

中图分类号：TP309　　**文献标志码：**A　　**文章编号：**1674-2605(2024)03-0007-06

DOI：10.3969/j.issn.1674-2605.2024.03.007

Network Information Recognition and Classification Algorithm Based on Matching Autonomous Learning

ZENG Guanghui
(Guangzhou Institute of Technology, Guangzhou 510900, China)

Abstract: To improve the accuracy of network information recognition and classification, a network information recognition and classification algorithm based on matching autonomous learning is proposed to address the high dimensionality, high noise and other characteristics of massive network information. Firstly, using support vector machine to recognize network information; Then, a retrieval matrix is constructed using singular value decomposition algorithm for singular value decomposition and correlation queries; Finally, calculate the similarity matching degree of network information, and input the network information with higher matching degree into the convolutional neural network for learning and training, outputting the network information classification results. The experimental results show that the network information recognition accuracy of the algorithm reaches over 97.90%, and the average classification accuracy for different types of network information is 98.04%, which has certain practical application effectiveness.

Keywords: matching autonomous learning; network information; support vector machine; singular value decomposition; convolutional neural network; recognition and classification

0 引言

在当前的信息时代，网络信息呈海量式与爆炸式增长[1]。网络信息不仅涉及多个特征，如文本内容、图像像素、格式类别等，还包含大量的干扰或噪声，如文本拼写错误、图像噪点或失真、网络攻击等，故其应用性与安全性受到相关研究人员的重视。网络信息的识别与分类是提升其应用性与安全性的基础[2-3]。

周家恺等[4]基于朴素贝叶斯对网络信息进行特征识别，识别效率较高，但易受来源数据噪声影响，识别精准度还有一定的提升空间。朱方娥等[5]提出基于分类规则挖掘的数据多标记特征分层识别方法，特征识别及分类的准确度较高，但较为依赖数据来源，需进行更加完善的数据预处理。

本文提出一种基于匹配自主学习的网络信息识别与分类算法。通过支持向量机、奇异值分解算法、卷

* 基金项目：2021 年度广东省教育厅教育科学规划课题（高等教育专项）“教育信息化 2.0 背景下高职 PAE 翻转课堂教学模式研究”（2021GXJK613）

积神经网络的匹配应用，实现网络信息的识别与分类。

1 算法流程

匹配自主学习算法是指对输入数据进行匹配并比较的自主学习算法。自主学习算法以多智能体深度强化学习类方法为代表，通过构建认知智能体，自动学习和获取复杂系统深层次的规律[6]。

本文引入自主学习算法中的支持向量机、奇异值分解算法、卷积神经网络对网络信息进行识别与分类。基于匹配自主学习的网络信息识别与分类算法流程如图 1 所示。

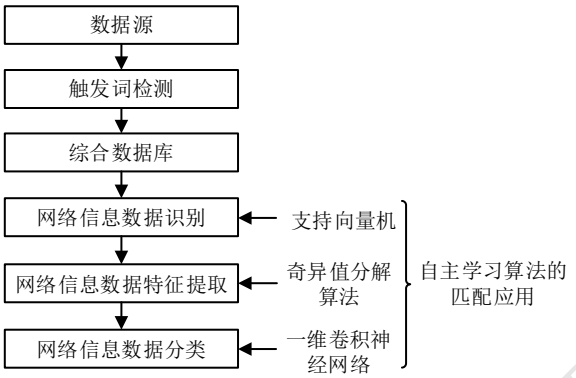

图 1　基于匹配自主学习的网络信息识别与分类算法流程

本文利用支持向量机实现网络信息的识别；采用奇异值分解算法提取异常网络信息的特征向量；运用一维卷积神经网络算法分类处理异常网络信息[7]。其中，一维卷积神经网络不易出现信息损耗、丢失和信息畸变等问题，可提升网络信息识别与分类的效率。

2 网络信息识别

支持向量机能够处理高维特征空间的分类问题，通过构建最优的超平面来实现数据分类，可有效地处理小样本问题，且对未见过的数据具有较好的泛化能力，减少过拟合风险。首先，对网络信息进行触发词检测预处理，并均等分为训练数据集与测试数据集[8]；然后，利用支持向量机对训练数据集中的数据进行训练，构建网络信息识别模型；接着，将测试数据输入到网络信息识别模型，界定网络信息识别阈值；最后，得到网络信息的识别结果。基于支持向量机的网络信息识别流程如图 2 所示。

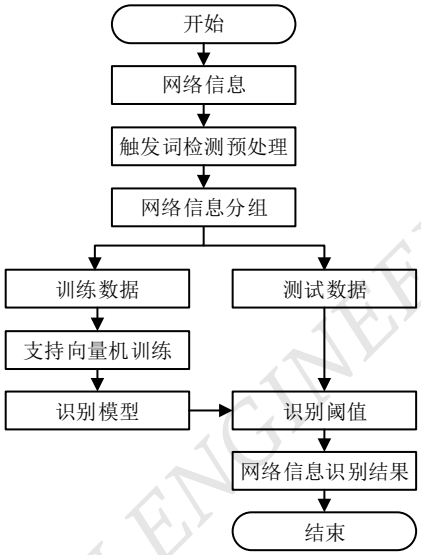

图 2　基于支持向量机的网络信息识别流程

网络信息的主要特征表现为海量性与高度开放性[9]。若对全部的网络信息都进行分析处理，将降低网络信息的识别效率，同时分析不重要的网络信息也会提升识别成本。因此在对网络信息进行识别与分类前，采用触发词检测方法对网络信息进行预处理，清除网络信息中的无用信息，减少数据维度。触发词是可最大限度地反映事件的词语。

利用支持向量机训练网络信息识别模型，获得网络信息性能指标曲线与阈值查阅表，由此可获取网络信息识别阈值[10]。基于该网络信息识别阈值对测试数据集的网络信息进行识别，确定其为正常网络信息或异常网络信息。基于支持向量机的网络信息识别，以径向基函数（radial basis function, RBF）为核函数，以接受者操作特性曲线（receiver operating characteristic curve, ROC）与坐标轴围成的面积（area under curve, AUC）为识别参数优化指标，对惩罚系数与核函数进行优化；同时引入交叉验证的方法避免支持向量机出现过拟合。

基于支持向量机的网络信息识别过程描述如下：

设 $T=\{(q_1,y_1),(q_2,y_2),\cdots,(q_i,y_i)\}$ 为训练数据集，$q_i\in R^n$ 和 $y_i\in\{+1,-1\}$ $(i=1,2,\cdots,k)$ 分

别表示第 i 个网络信息向量和 q_i 的类别标签。以 (q_i, y_i) 为样本点，构建最优超平面，即网络信息识别模型 N 的计算过程为

$$N = \min \frac{\|\boldsymbol{w}\|^2}{2} + C\sum_{i=1}^{k}\xi_i \tag{1}$$

式中：$\boldsymbol{w}$ 为超平面的法向量，C 和 ξ_i 分别为惩罚系数与网络信息识别的误项。

其约束条件为

$$\text{s.t.}\begin{cases} y_i\left(\boldsymbol{w}^{\mathrm{T}} q_i + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \tag{2}$$

式中：b 为超平面的常数项。

3 网络信息分类

3.1 基于奇异值分解算法的相似性匹配度计算

在网络信息特征分类之前，先对网络信息进行特征提取[11]。将网络信息识别模型输出的识别结果作为输入，采用奇异值分解算法进行特征提取。通过隐含语义提取，清除不相关词汇，得到关键词向量，目标特征描述矩阵向量间的内在属性；对目标特征进行变换与分解处理，得到的相似性匹配结果作为输出，具体过程如下：

1） 构建词条——文档矩阵，对待提取特征的网络信息文档进行处理，清除不相关词汇，获取网络信息文档的关键词向量，维数为 n。若网络信息文档包含 m 个文件，则可获取一个 $n \times m$ 维矩阵。奇异值分解算法将词条——文档矩阵分解为 3 个不同的矩阵，公式描述为

$$\boldsymbol{X} = \frac{\boldsymbol{R}\Sigma\boldsymbol{G}^{\mathrm{T}}}{N} \tag{3}$$

式中：$\boldsymbol{R}$ 描述网络信息文档内不同词条间的相关性[12]，$\boldsymbol{G}$ 描述网络信息不同文档间的相关性，$\boldsymbol{R}$ 与 $\boldsymbol{G}$ 均为正交矩阵；Σ 为对角矩阵。

考虑到矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{G}$ 均具有线性独立特性，可通过近似矩阵 $\boldsymbol{X}_K$ 取代 $\boldsymbol{X}$ 进行分析，如公式(4)所示。

$$\boldsymbol{X}_K = \frac{\boldsymbol{R}_K\Sigma_K\boldsymbol{G}_K^{\mathrm{T}}}{N} \tag{4}$$

式中：$\boldsymbol{R}_K$ 和 $\boldsymbol{G}_K$ 分别为 $\boldsymbol{R}$ 和 $\boldsymbol{G}$ 的前 K 列，Σ_K 为包含 $\boldsymbol{X}$ 的前 K 个最大奇异值，$\boldsymbol{R}_K^{\mathrm{T}}\boldsymbol{R}_K = 1$；$\boldsymbol{G}_K^{\mathrm{T}}\boldsymbol{G}_K = 1$，由此可提升特征提取效率。

2） 网络信息文档中的若干个关键词通过变换生成一个 K 维向量 $\boldsymbol{Q}_P$，其代表一个虚文档，将 $\boldsymbol{Q}_P$ 与文档相关性矩阵 $\boldsymbol{G}$ 内的文档向量进行对比，得到相似性匹配结果[13] $\boldsymbol{Q}_P$ 的计算公式为

$$\boldsymbol{Q}_P = \boldsymbol{X}_K\boldsymbol{R}\Sigma^{-1} \tag{5}$$

$\boldsymbol{Q}_P$ 值越大，表明网络信息相似性匹配度越高，分类效果较好；$\boldsymbol{Q}_P$ 值越小，表明网络信息相似性匹配度越低，分类效果较差。

应用奇异值分解算法进行网络信息特征提取的过程中，在一定程度上去除了网络信息中的噪声或异常点，通过保留主要的奇异值和特征向量，可以恢复经去噪处理后的原始信息，为后续网络信息的精准分类提供保障。

3.2 基于卷积神经网络的网络信息分类

根据网络信息特征建立一维卷积神经网络分类模型，实现网络信息的分类处理。一维卷积神经网络模型包括输入层、卷积层、池化层、全连接层、输出层等，输出层可输出网络信息的分类结果，结构如图 3 所示。

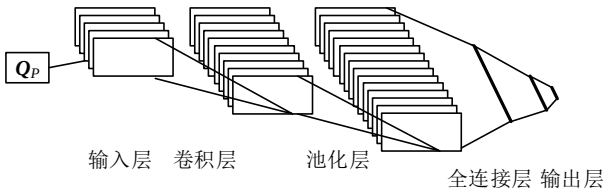

图 3 一维卷积神经网络模型结构

一维卷积神经网络模型的输入是相似性匹配结果。

卷积层作为一维卷积神经网络的核心，主要负责对网络信息进行稀疏连接[14]，降低网络信息特征的参数量。利用公式(6)确定卷积层的输出 c_i 为

$$c_i = \sigma \boldsymbol{Q}_P \cdot \left(\sum_{j=1}^{e} T_i^j + YZ \right) \tag{6}$$

式中：σ 和 e 分别为激活函数和网络信息特征数量，T_i^j 为第 j 个网络信息数据，Y 和 Z 分别为输出偏置与卷积核尺寸。

池化层主要负责进一步降低卷积层输出的特征参数，同时保留网络信息的主要特征[15]。利用公式(7)描述最大池化函数为

$$p_i^j = \max\left(c_i + hf\right) \tag{7}$$

式中：h 和 f 分别为池化层移动步长和池化尺寸。

在卷积层与池化层的逐渐堆叠下，不仅能够提取网络信息的深层特征，还能够显著降低参数量。

将提取的网络信息特征转换为一维向量 $p = \left(p_1, \cdots, p_I\right)$，并输出至全连接层进行分类，其中 I 为最后一层池化层的神经节点数量。

输出层是一维卷积神经网络的最后一层，其输出的结果即为网络信息所属类别 U_k：

$$U_k = \frac{\exp\left(z_k\right)}{\sum_{k=1}^{N_c} \exp\left(z_k\right)} \cdot p_i^j \tag{8}$$

式中：z_k 为全连接层的输出，k 和 N_c 分别为网络信息类别的索引和全部网络信息的数量。

4 实验分析

4.1 实验准备

为验证本文算法在实际网络信息识别与分类中的效果，分别对网络信息识别、特征向量提取、网络信息分类的性能进行测试。

实验环境为 Ubuntu 18.04 操作系统，Python3.7 编程语言，TensorFlow2.0 开发框架，具备 GPU 加速功能的 NVIDIA GeForce RTX 2080 Ti。计算资源方面，Intel Core i7-8700K CPU @ 3.70 GHz 的计算机，32 GB 内存。触发词匹配阈值设定为 0.7，当网络信息中某个词与触发词的相似度高于 0.7 时，将该词输入到综合数据库中以待后续处理；当网络信息中某个词与触发词的相似度低于 0.7 时，忽略或丢弃该词。奇异值分解降维维度设置为 100 维，卷积核大小设置为 3、5 和 7，以便对不同尺度的网络信息进行特征提取，利用最大池化对网络信息特征进行降维。

实验数据集选取 KDD cup 99 数据集，包括正常网络信息（文本信息、图片信息、视频信息）、异常网络信息（虚假信息、攻击信息）共 4 909 542 条。其中，攻击信息包含 6 种类型，如表 1 所示。

表 1　6 种类型攻击信息描述

编号	攻击名称	攻击描述
1	DOS 攻击	拒绝服务攻击，令计算机或网络无法提供正常服务
2	MSCI 攻击	恶意状态命令注入攻击
3	跨站点脚本攻击	页面加载过程中，浏览器上履行恶意脚本
4	NMRI 攻击	简单的恶意响应注入攻击
5	SQL 注入攻击	在本文框内输入 SQL 语句用于拐骗应用程序显现
6	Reconnaissance 攻击	侦察攻击

4.2 支持向量机训练

选取 KDD cup 99 数据集的 50%，即 2 454 771 条网络信息作为训练样本进行训练，得到支持向量机的网络信息识别准确率波动图如图 4 所示。

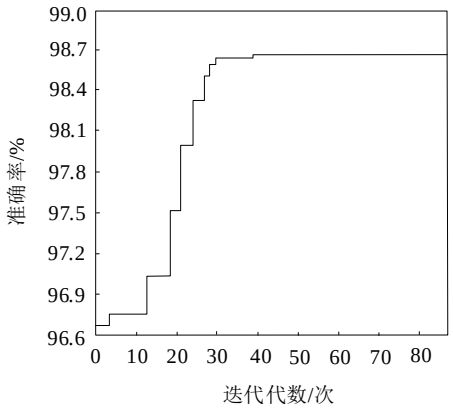

图 4　支持向量机的网络信息识别准确率波动图

由图 4 可知：随着支持向量机迭代次数的增加，网络信息识别准确率也逐渐提高，当迭代次数小于 30 次时，识别准确率提高速度较快；当迭代次数大于 30 次时，识别准确率提高速度逐渐趋于平缓；当迭代次

数达到 40 次时，识别准确率稳定在 98.70%左右。至此，完成网络信息识别模型的训练。

4.3 实验结果与分析

4.3.1 网络信息识别性能测试

利用训练好的网络信息识别模型对 KDD cup 99 数据集剩余的 50%，即 2 454 771 条网络信息进行识别，判断网络信息状态。为验证本文算法的性能，选取文献[4]的基于朴素贝叶斯方法和文献[5]的基于分类规则挖掘方法进行对比实验，结果如表 2 所示。

表 2 3 种方法的测试样本识别准确率

测试样本数据量/条	本文算法	基于朴素贝叶斯方法	基于分类规则挖掘方法
10 000	99.21%	97.87%	95.79%
100 000	98.73%	96.92%	95.06%
1 000 000	98.19%	96.48%	94.63%
2 454 771	97.90%	95.73%	93.57%

由表 2 可知：随着测试样本数据量的增加，3 种方法的识别准确率均有所下降，在测试样本数据量为 10 000 条时，本文算法、基于朴素贝叶斯方法、基于分类规则挖掘方法的识别准确率最高，分别为 99.21%、97.87%、95.79%；在测试样本数据量为 2 454 771 条时，3 种方法的识别准确率最低，分别为 97.90%、95.73%、93.57%，表明测试样本数据量对准确率造成影响，且本文算法具有较高的网络信息识别性能。

4.3.2 特征向量提取性能测试

对于相同的网络信息，不同方法提取的特征向量会有所差异。采用本文算法与基于朴素贝叶斯方法、基于分类规则挖掘方法分别进行网络信息特征向量提取性能对比实验，以方差$V(X)$、偏度$S(X)$、峰度$K(X)$为评估指标。其中，方差越大，说明样本数据在这一维度上的差异性越大，数据包含大量的噪声或异常值；偏度用于衡量数据分布的不对称性，正值表示数据右偏，负值表示数据左偏，接近 0 表示数据近似对称；峰度正值表示尖峭峰，即比正态分布更集中，而负值表示平坦峰，即比正态分布更平缓，峰值大说明存在极端值。3 种评估指标的计算公式为

$$\begin{cases} V(X)=\dfrac{1}{n}\sum\left(x_i-\mu\right)^2 \\ S(X)=\dfrac{1}{n}\sum\dfrac{\left(x_i-\mu\right)^3}{\sigma^3} \\ K(X)=\dfrac{1}{n}\sum\dfrac{\left(x_i-\mu\right)^4}{\sigma^4} \end{cases} \tag{9}$$

式中：X 为整体的样本数据，n 为样本数量，x_i 为第 i 个样本数据，μ 为样本均值。

本文算法与基于朴素贝叶斯方法、基于分类规则挖掘方法的特征向量提取性能对比结果如表 3 所示。

表 3 特征向量提取性能对比结果

关键词数量	本文算法			基于朴素贝叶斯方法			基于分类规则挖掘方法		
	方差	偏度	峰度	方差	偏度	峰度	方差	偏度	峰度
3	0.87	−0.01	0.92	0.94	0.07	0.95	1.02	−0.05	1.00
5	0.89	−0.03	0.94	0.96	0.03	0.97	1.03	−0.05	1.02
10	0.91	−0.06	0.90	0.96	0.02	0.96	1.51	−0.10	1.03
20	0.95	−0.05	0.88	0.98	0.08	0.95	2.04	−0.12	1.04

由表 3 可知：随着关键词数量逐渐增加，3 种方法的特征向量方差也逐渐增大，说明关键词数量越多，特征提取的难度越大，越容易存在噪声；在不同关键词数量下，3 种方法均保持较小且接近 0 的偏度，表明特征向量分布相对对称，本文算法的偏度值稳定且偏负，显示特征向量分布可能略向左偏，相比之下，另外 2 种方法在关键词数量增多时偏度值增加，说明其分布偏斜较大；本文算法特征向量的峰度相对另外 2 种方法较低，说明特征提取后，极端值较少，特征向量提取效果较好。

4.3.3 网络信息分类性能测试

采用本文算法、基于朴素贝叶斯方法、基于分类规则挖掘方法对识别的 2 454 771 条网络信息进行分类处理，结果如表 4 所示。

表 4　网络信息分类处理结果

状态	类别	实际样本数/条	分类处理准确率/%		
			本文算法	基于朴素贝叶斯方法	基于分类规则挖掘方法
正常	文本信息	2 083 234	96.82	93.32	90.12
	图片信息	243 720	97.90	94.78	91.81
	视频信息	94 975	98.05	95.12	92.33
异常	虚假信息	9 537	98.87	95.63	92.02
	DOS 攻击	3 425	98.54	95.22	93.09
	MSCI 攻击	3 687	98.23	96.87	93.75
	跨站点脚本攻击	4 139	97.96	95.93	92.23
	NMRI 攻击	3 972	97.71	95.75	93.98
	SQL 注入攻击	3 539	98.88	95.02	92.16
	Reconnaissance 攻击	4 543	97.46	95.25	92.86

由表 4 可知：本文算法的平均分类准确率为 98.04%；基于朴素贝叶斯方法和基于分类规则挖掘方法的平均分类准确率分别为 95.29%和 92.44%，验证了本文算法对网络信息的分类准确率较高、分类处理性能较好。对异常网络信息的精准分类能够更好地对攻击信息采取相应的防御措施。

5　结论

本文研究基于匹配自主学习的网络信息识别与分类算法，利用自主学习算法中的支持向量机、奇异值分解算法、一维卷积神经网络实现网络信息的识别与分类。实验结果显示，该算法的网络信息识别准确率、特征向量提取性能以及网络信息分类准确率均较高，说明该算法具有较好的应用性能。在本文算法研究的过程中，受时间与经费的限制，在处理大规模网络信息时，算法的运行效率受到一定程度的限制。因此，未来将会探索更高效和可扩展的算法形式，以应对大规模网络信息的识别与分类。

参考文献

[1] 周毅,张雪.网络信息内容生态安全风险整体智治的理论框架与实现策略研究[J].图书情报工作,2022,66(5):44-52.

[2] 韩正彪,马毛宁,翟冉冉.网络学术信息搜索中用户情感的识别及变化研究[J].情报学报,2022,41(3):314-324.

[3] 蒋岑,吴迪.隐蔽无线通信网络传输信息云存储密文检索[J].计算机仿真,2021,38(6):125-128;137.

[4] 周家恺,綦方中.网络流量时延特征数据的识别方法仿真[J].计算机仿真,2022,39(5):398-401;460.

[5] 朱方娥,郭建方,曹丽娜.基于分类规则挖掘的数据多标记特征分层识别[J].计算机仿真,2021,38(4):310-314.

[6] 朱晓慧,钱丽萍,傅伟.基于生成对抗网络增强恶意代码的方法[J].计算机工程与设计,2021,42(11):3034-3042.

[7] 高昂,郭齐胜,董志明,等.基于 EAS+MAD RL 的多无人车体系效能评估方法研究[J].系统工程与电子技术,2021,43(12):3643-3651.

[8] 蒋丽,黄仕建,严文娟.基于低秩行为信息和多尺度卷积神经网络的人体行为识别方法[J].计算机应用,2021,41(3):721-726.

[9] 陆晓松,王国庆,李勖之,等.场地环境大数据采集和机器学习方法在污染智能识别中的应用研究进展[J].生态与农村环境学报,2022,38(9):1101-1111.

[10] 张泽锋,毛存礼,余正涛,等.融入领域术语词典的司法舆情敏感信息识别[J].中文信息学报,2022,36(9):76-83;92.

[11] 陈思佳,罗志增.基于长短时记忆和卷积神经网络的手势肌电识别研究[J].仪器仪表学报,2021,42(2):162-170.

[12] 向志华,梁玉英.基于机器学习的视频识别与自适应推送算法[J].沈阳工业大学学报,2022,44(3):336-340.

[13] 张玲,卫传征,林臻彪,等.一种基于机器学习的 Tor 网络识别探测技术[J].电子技术应用,2021,47(4):54-58.

[14] 华萌萌,尹君,胡召玲,等.基于机器学习的历史气候重建论文智能识别与数据挖掘初探[J].第四纪研究,2021,41(2):550-561.

[15] 宋雅文,杨志豪,罗凌,等.基于字符卷积神经网络的生物医学变异实体识别方法[J].中文信息学报,2021,35(5):63-69.

作者简介：

曾光辉，男，1972 年生，本科，副教授，主要研究方向：智能信息处理。E-mail: cupid74@163.com