基于邮件头的异常邮件检测

张 毓 杨 涛 姬杨帆 顾知义
（内蒙古大学）

DOI：10.65285/ZRKXLT.V1I1.004

摘 要：目的是验证仅利用邮件头信息检测垃圾邮件和钓鱼邮件的可行性，并比较监督与无监督模型性能。方法 基于 TREC 2007 和 Monkey.org 数据集，从 76707 封邮件中提取缺失字段、字符串模式、Received 头一致性、域名匹配等 94 项特征，构建 10 种监督模型、单类支持向量机和孤立森林，以准确率、精确率、召回率、F1 值和 AUC 评价性能。结果 垃圾邮件任务中，堆叠集成准确率和 F1 值分别为 .9971 和 .9978，随机森林分别为 .9966 和 .9975；多种监督模型在钓鱼邮件任务中的准确率达到 1.0000。无监督模型中，孤立森林检测垃圾邮件的准确率为 .7592，单类支持向量机检测钓鱼邮件的准确率为 .7403。域名一致性和关键头字段缺失是主要判别特征。结论 邮件头特征能够有效支持异常邮件检测，监督模型整体优于无监督模型，但仍需通过跨来源、跨时间数据验证泛化能力。

关键词：异常邮件检测；邮件头；机器学习；集成学习；无监督学习

Abnormal Email Detection Based on Email Headers

Yu Zhang, Tao Yang, Yangfan Ji, Zhiyi Gu
Inner Mongolia University

Abstract: This study aims to verify the feasibility of identifying spam and phishing emails merely through email header information, and compare the performance of supervised and unsupervised detection models. Based on the TREC 2007 and Monkey.org datasets, a total of 94 features including missing header fields, string patterns, Received header consistency and domain name matching are extracted from 76707 email samples. Ten supervised machine learning models, as well as one-class SVM and isolation forest unsupervised models are constructed. Model performance is evaluated in terms of accuracy, precision, recall, F1-score and AUC. The results show that in spam detection tasks, the stacking ensemble model achieves an accuracy of 0.9971 and an F1-score of 0.9978, while the random forest model reaches 0.9966 and 0.9975 respectively. Multiple supervised models obtain an accuracy of 1.0000 in phishing email detection. For unsupervised models, the isolation forest achieves an accuracy of 0.7592 for spam detection, and the one-class SVM achieves an accuracy of 0.7403 for phishing detection. Domain name consistency and the absence of key header fields are the core discriminative features for abnormal emails. Conclusion: Email header features can support effective abnormal email detection. Supervised models outperform unsupervised models overall. Further cross-source and cross-temporal data validation is required to verify the generalization performance of the models.

Keywords: abnormal email detection; email header; machine learning; ensemble learning; unsupervised learning

前言：

垃圾邮件和钓鱼邮件会造成账号泄露、隐私损害、经济损失及品牌风险。反钓鱼工作组报告显示，2022 年第 4 季度共观测到 1350037 次钓鱼攻击，异常邮件仍是网络安全治理中的持续性问题 [1]。传统过滤方法多依赖主题、正文词频或语义内容，虽然能够获取丰富信息，但其性能可能受到语言、文本编码和内容规避策略影响，同时还涉及正文隐私及较高的特征处理成本。

邮件头记录 From、Sender、Reply-To、Received 和 Message-ID 等传输与身份信息，可判断字段是否缺失、发件域与退信域是否一致及传输路径是否异常。Beaman 和 Isah 提出仅利用邮件头特征检测垃圾邮件与钓鱼邮件并同时比较多类监督模型和单类异常检测模型 [2]。该思路不读取正文，特点是维度较低、语言依赖较弱和便于在线处理。

本研究在保留原始实验数据的基础上，将任务整理为规范的实证研究：从两类公开邮件数据中提取 94 项邮件头特征，构建 10 种监督学习模型及 2 种无监督模型，比较垃圾邮件与钓鱼邮件检测性能，并通过置换重要性分析主要判别信号。

一、方法

（一）数据来源

实验使用 TREC 2007 垃圾邮件数据和 Monkey.org 钓鱼邮件数据。TREC 2007 数据由 25220 封正常邮件和 50199 封垃圾邮件组成，共 75419 封；该数据集建立了按时间顺序进行垃圾邮件过滤与人工金标准比较的评测框架 [3]。Monkey.org 数据包含 1288 封钓鱼邮件。垃圾邮件任务使用

TREC 2007 中 Spam 与 Ham，钓鱼邮件任务将 Monkey.org 中的 Phishing 与 TREC 2007 中的 Ham 组合，共得到 26508 个样本。

（二）邮件头特征提取

原始邮件经解析后提取 94 项特征，分为 6 类：45 项缺失头字段特征、11 项字符串或编码特征、11 项数值特征、4 项 Received 头一致性特征、22 项域名匹配特征和 1 项跳数特征。特征提取流程为：解析原始邮件，读取标准与扩展头字段，计算缺失标记、字符串模式、数值统计和域名一致性，最后形成邮件—特征矩阵。垃圾邮件样本多于正常邮件，钓鱼邮件数量最少；不同类别在跳数、缺失字段数量和域名匹配得分上呈现可见差异，为后续分类提供了基础。

（三）实验设计与模型

实验在 Windows 10 与 Python 3.8 以上环境完成，机器学习模型由 scikit-learn 实现 [4]。监督学习任务采用 25% 的训练集和 75% 的测试集。垃圾邮件与钓鱼邮件分类均使用 45 项精简特征并通过 StandardScaler 依据训练集参数完成标准化。无监督实验使用 61 项扩展特征，只以正常邮件拟合模型，再在包含正常与异常邮件的测试集上评价。

监督学习模型包括随机森林、多层感知机、逻辑回归、支持向量机、决策树、高斯朴素贝叶斯、AdaBoost、梯度提升、K 近邻和堆叠集成。随机森林通过多棵决策树投票降低单树方差 [5]；堆叠集成以随机森林、多层感知机、K 近邻和支持向量机为基础学习器，以逻辑回归为元学习器，通过组合不同分类边界提高稳定性 [6]。无监督部分采用单类支持向量机和孤立森林。单类支持向量机根据正常样本估计高维分布的支持区域，将边界外样本视为异常 [7]；孤立森林通过随机选择特征和切分点隔离样本，异常样本通常具有较短的平均路径 [8]。两种模型均只使用正常邮件训练。

（四）评价指标与解释方法

以异常邮件为正类，根据混淆矩阵计算准确率、精确率、召回率和 F1 值，并绘制 ROC 曲线、计算 AUC。准确率衡量总体正确率，精确率与召回率分别反映误报和漏报情况，F1 值综合两者；无监督任务采用相同指标。

采用置换重要性解释模型：随机打乱单项特征并观察性能下降，降幅越大表明模型对该特征依赖越强。该指标不代表因果关系。

二、结果

（一）垃圾邮件监督学习结果

表 2 显示，堆叠集成的准确率和 F1 值最高，分别为 .9971 和 .9978；随机森林分别为 .9966 和 .9975。其余非线性模型准确率均超过 .9940；高斯朴素贝叶斯准确率最低（.9212），但召回率为 .9642，呈现较高检出率与较多误报并存的特点。

表 1　垃圾邮件监督学习性能

排名	模型	Accuracy	Precision	Recall	F1-Score
1	Stacking Ensemble	0.9971	0.9981	0.9975	0.9978
2	Random Forest	0.9966	0.9976	0.9973	0.9975
3	MLP	0.9959	0.9978	0.9960	0.9969
4	KNN	0.9955	0.9979	0.9954	0.9966
5	SVM	0.9952	0.9980	0.9948	0.9964
6	Gradient Boosting	0.9947	0.9962	0.9959	0.9960
7	Decision Tree	0.9931	0.9962	0.9935	0.9948
8	Logistic Regression	0.9627	0.9683	0.9759	0.9721
9	AdaBoost	0.9572	0.9589	0.9777	0.9682
10	Gaussian NB	0.9212	0.9211	0.9642	0.9422

（二）钓鱼邮件监督学习结果

在 Phishing 与 Ham 任务中，随机森林、支持向量机、高斯朴素贝叶斯和堆叠集成的准确率与 F1 值均为 1.0000，其他模型准确率也超过 .9980。结果表明当前两类样本在域名一致性和字段完整性方面差异明显。

（三）无监督学习结果

仅使用正常邮件训练时，孤立森林检测垃圾邮件的准确率和 F1 值分别为 .7592 和 .7447；单类支持向量机检测钓鱼邮件分别为 .7403 和 .7430。其余组合表现较低，说明单类模型的效果受正常样本覆盖范围及异常类型影响（表 2）。

表 2　无监督异常检测性能

模型	异常类型	Accuracy	Precision	Recall	F1-Score
Isolation Forest	Spam	0.7592	0.7924	0.7025	0.7447
One-Class SVM	Phishing	0.7403	0.7354	0.7508	0.7430
Isolation Forest	Phishing	0.6079	0.5930	0.6879	0.6370
One-Class SVM	Spam	0.4204	0.4520	0.7504	0.5642

图 1 显示，监督堆叠集成的准确率和 F1 值均接近 1，明显优于无监督模型。无监督方法无需异常标签，更适合作为未知威胁初筛或二级告警机制。

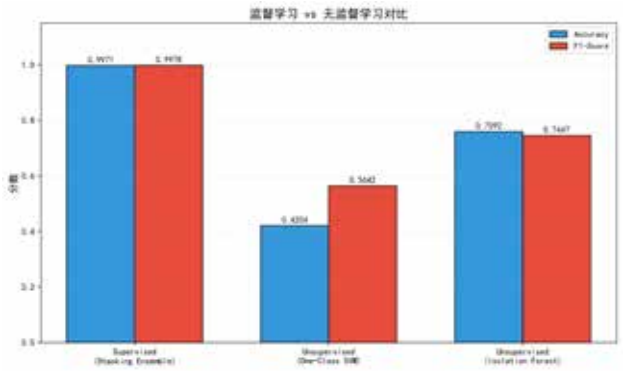

图 1 监督学习与无监督学习性能比较

（四）特征重要性结果

在置换重要性排名中 domain_match_errors-to_sender、missing_x-beenthere 和 missing_list-archive 的重要性分数均为 .280，missing_errors-to 为 .275，missing_sender 为 .251。前 10 项特征中，缺失字段和域名匹配特征占据多数，说明模型主要利用邮件头完整性及身份字段之间的一致关系，而非邮件正文语义完成判别。

三、讨论

（一）监督模型性能差异

监督模型整体性能较高，说明邮件头特征能够稳定表示当前数据中的异常模式。堆叠集成综合树模型、神经网络、核方法和近邻方法，取得最高 F1 值；随机森林略低，但训练和维护成本更低，具有较好的性能—成本平衡。逻辑回归受线性边界限制，高斯朴素贝叶斯的条件独立假设难以表达缺失字段、域名匹配和传输路径之间的关联。Gaussian NB 召回率较高但精确率较低，也表明模型选择应同时考虑误报与漏报成本。

（二）无监督模型的作用

无监督模型与监督模型之间存在明显性能差距，原因在于单类方法只学习正常邮件分布，无法直接利用已知垃圾邮件或钓鱼邮件的判别信息。尽管如此，无监督学习仍具有现实价值。面对新型攻击或缺少高质量标签的组织，单类支持向量机和孤立森林可以先对偏离正常分布的邮件进行排序，再交由规则系统、监督模型或安全人员复核。孤立森林在垃圾邮件上的 F1 值达到 .7447，表明其具备一定初筛能力。后续可通过滑动时间窗口更新正常样本、按部门或邮件系统建立分层基线，并根据告警容量调整异常阈值。

结语：

本研究从 76707 封邮件中提取 94 项邮件头特征，比较 10 种监督模型和 2 种无监督模型。堆叠集成和随机森林在垃圾邮件检测中表现最佳，多种监督模型在当前钓鱼邮件数据上实现完全分类；无监督模型无需异常标签，但性能较低。域名一致性和关键字段缺失是主要判别信号。邮件头检测具有不读取正文和语言依赖较弱的优势，其跨来源泛化能力仍需外部验证。

参考文献：

[1] Anti-Phishing Working Group. Phishing Activity Trends Report, 4th Quarter 2022[EB/OL]. https://docs.apwg.org/reports/apwg_trends_report_q4_2022.pdf, 2023-05-10.

[2] Beaman, C. and Isah, H. (2022) Anomaly Detection in Emails Using Machine Learning and Header Information[EB/OL]. https://arxiv.org/abs/2203.10408, 2022-03-19.

[3] Cormack, G.V. (2007) TREC 2007 Spam Track Overview. In: Voorhees, E.M. and Buckland, L.P., Eds., Proceedings of the Sixteenth Text REtrieval Conference, National Institute of Standards and Technology, Gaithersburg.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011) Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[5] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

[6] Wolpert, D.H. (1992) Stacked Generalization. Neural Networks, 5, 241-259.

[7] Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J. and Williamson, R.C. (2001) Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13, 1443-1471.

[8] Liu, F.T., Ting, K.M. and Zhou, Z.H. (2008) Isolation Forest. Proceedings of the 2008 IEEE International Conference on Data Mining, IEEE, Pisa, 413-422.

作者简介：

顾知义（2001.10-）、男、汉族、籍贯江苏南京、硕士研究生在读、研究方向：计算机视觉、人工智能。

杨涛（2001.08-）、男、汉族、籍贯河南商丘、硕士研究生在读、研究方向：计算机视觉、人工智能。

张毓（2002.12-）、男、汉族、内蒙古呼和浩特市、硕士研究生在读、研究方向：多智能体长文档写作。

姬杨帆（2002.09-）、男、汉族、山西临汾、硕士研究生、主要研究方向：计算机视觉、人工智能。