**Software Engineering and Applications** 软件工程与应用, 2021, 10(2), 149-155
Published Online April 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.102017

Hans汉斯

# 面向网络安全的数据融合技术研究

**张海霞[1*]，吴建英[2]，黄克振[1]，连一峰[1]**

[1]中国科学院软件研究所，可信计算与信息保障实验室，北京
[2]北京市公安局网络安全保卫总队，北京
Email: *zhanghx@tca.iscas.ac.cn

收稿日期：2021年3月12日；录用日期：2021年4月8日；发布日期：2021年4月15日

## 摘　要

**网络安全保护工作需要对各类多源异构的网络安全数据进行融合分析，以提炼知识线索，发现数据内容之间的互补关系、隐含关系和关联关系，从而支撑网络安全监测发现、分析研判和处置应对等工作内容。本文从融合对象、融合目标和融合方法入手，提出了网络安全数据融合要素，重点阐述了基于统计学、数据挖掘和人工智能的数据融合方法，并对其应用于网络安全领域的适用性进行了探讨。**

## 关键词

**网络安全，数据融合，数据治理，数据挖掘，人工智能**

# Research on Technologies of Data Fusion for Network Security

**Haixia Zhang[1*], Jianying Wu[2], Kezhen Huang[1], Yifeng Lian[1]**

[1]Trusted Computing and Information Assurance Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing
[2]Network Security Corps of Beijing Municipal Public Security Bureau, Beijing
Email: *zhanghx@tca.iscas.ac.cn

Received: Mar. 12th, 2021; accepted: Apr. 8th, 2021; published: Apr. 15th, 2021

## Abstract

**Network security protection requires the fusion and analysis of all kinds of multi-source and heterogeneous data, so as to extract knowledge clues and discover the implicit correlation relation-**

*通讯作者。

文章引用: 张海霞, 吴建英, 黄克振, 连一峰. 面向网络安全的数据融合技术研究[J]. 软件工程与应用, 2021, 10(2): 149-155. DOI: 10.12677/sea.2021.102017

**ship between data items, and furthermore, to support the duties of security monitoring, research and judgement, disposal and response. Starting with objects, targets and methods of data fusion, this paper puts forward the elements of network security data fusion, focuses on the data fusion methods based on statistics, data mining and artificial intelligence, and discusses the applicability in the field of network security.**

## Keywords

**Network Security, Data Fusion, Data Governance, Data Mining, Artificial Intelligence**



## 1. 数据融合技术概述

数据融合[1]的概念是 20 世纪 70 年代提出来的，基本原理就像人脑综合处理信息一样，充分利用多个传感器资源，通过对多传感器及其观测信息的合理支配和使用，把多传感器在空间和时间上可冗余或互补的信息，依据某种规则进行组合，以获得被测对象的一致性解释。在 20 世纪 80 年代，数据融合技术引起了世界范围内的广泛关注，美国等西方国家陆续部署实施了围绕数据融合的重大研究项目，取得了一系列关键技术的突破。

围绕数据融合技术的应用实践也在各个社会领域广泛开展。特别是随着大数据时代的来临，制造业、建筑业、零售业、交通运输业、社会服务业、军事领域、遥感遥测领域、医疗领域、环保领域、城市规划领域、资源管理领域等各行各业均迎来了对数据融合技术的应用需求，迫切需要对多源、异构、海量的原始数据、基础数据和业务数据进行规模化、智能化的融合处理，实现业务驱动下的数据归并、组合、串联、拼接，以获得对业务目标的完整刻画，支撑业务研判决策。

利用多个传感器数据，获取关于对象和环境的全面、完整信息，其核心问题是选择合适的融合算法。多传感器系统的数据具有多样性和复杂性，数据融合算法的基本要求是具有鲁棒性和并行处理能力。此外还有运算速度、精度、接口、适配能力等要求。一般情况下，基于非线性的数学方法，如果具有较好的容错性、自适应性、联想记忆能力和并行处理能力，则都可以用来作为数据融合方法。

数据融合虽然并未形成独立的理论体系，但在不同领域根据各自的具体应用背景，已经提出了许多成熟并且有效的融合方法。数据融合的常用方法基本上可概括为随机类算法和人工智能算法两大类，随机类算法有加权平均法、卡尔曼滤波法、多贝叶斯估计法、Dempster-Shafer (D-S)证据推理、产生式规则等；人工智能类则有模糊逻辑理论、神经网络、粗集理论、专家系统等。可以预见，神经网络和人工智能等新概念、新技术在数据融合中将起到越来越重要的作用。

本文将基于网络安全多源异构数据汇聚与融合的实践，对网络安全数据融合涉及的要素进行划分，接着对适用于网络安全领域的三类融合技术进行分析对比，最后对网络安全数据融合技术进行归纳总结。

## 2. 网络安全数据融合要素

当前，网络空间安全形势日益严峻，为了有效保护关键信息基础设施免受攻击、侵入、干扰和破坏，维护网络空间安全和秩序，应对来自国内外跨空间、跨领域的网络安全威胁，需要大力开展网络安全技术研究，并建立协同联动机制，实现针对网络威胁的常态化防御和体系化防御。

随着金融、交通、能源、公共服务等领域的信息化发展，网络空间逐渐延伸到我们的物理空间和社会空间。云计算平台、物联网、工业控制系统、移动互联网成为网络空间的重要组成部分。网络安全保护工作需要掌握网络资产、威胁和脆弱性等各类安全要素信息，对各类多源异构的网络安全数据进行数据融合，以发现数据内容之间的互补关系、隐含关系和关联关系，实现大数据支撑下的网络安全监测发现、分析研判和处置应对。本文重点对网络安全数据融合技术进行研究。

网络安全的数据融合有多个维度，我们将融合对象、融合目标和融合方法作为一级要素，在此基础上进行要素细化，提出如表 1 所示的数据融合要素清单，并针对每类要素进行示例说明。

**Table 1.** Data fusion elements of network security
**表 1.** 网络安全数据融合要素

| 一级要素 | 二级要素 | 三级要素 |
|---|---|---|
| 融合对象 | 流量数据 | 网络中传输的原始流量或经协议还原后的流量协议日志 |
| | 告警数据 | 监测发现的网络攻击、病毒、木马等安全告警信息 |
| | 日志数据 | 设备产生的各类操作系统日志、数据库日志、应用系统日志、安全日志 |
| | 网络资产数据 | 服务器、网络设备、存储设备、安全设备、数据库、终端等各类资产的详细信息 |
| | 网络架构 | 网络中各类设备、资产的拓扑结构、连接关系和信任关系 |
| | 基础知识数据 | IP 定位数据、域名注册信息、邮箱注册信息等 |
| | 安全知识数据 | 安全漏洞信息、木马病毒信息、补丁信息等 |
| | 威胁情报数据 | 恶意域名、恶意 IP 地址、恶意代码 MD5 等 |
| 融合目标 | 去重归并 | 将相同来源、相同目标或相同类型的安全事件进行合并 |
| | 数据过滤 | 将错误的告警数据进行过滤，避免对分析过程造成误导 |
| | 数据补全 | 对原始数据中缺失的数据项进行补充完善 |
| | 数据关联 | 分析挖掘多源数据中存在的关联关系 |
| | 时序关联 | 将不同时间段发生的相同来源或相同目标的攻击关联起来 |
| | 来源关联 | 将来自相同来源的攻击活动关联起来 |
| 融合方法 | 目标关联 | 将来自不同来源并针对相同目标的攻击活动关联起来 |
| | 因果关联 | 将攻击者实施的多个攻击步骤关联起来，还原攻击链 |
| | 聚类分析 | 利用聚类算法将正常网络行为与异常访问行为予以区分，以发现可疑攻击活动 |
| | 人工智能算法 | 分析挖掘数据中未知的隐含关系，为发现未知漏洞、新型攻击活动、新型木马病毒提供支撑 |

开展数据融合的网络安全数据来自于网络中的各类计算设备、存储设备、安全设备、安全系统或外部数据源，如防火墙、入侵检测系统、认证系统、流量探针、系统日志以及外部威胁情报信息，或采用资产探测技术感知网络中的资产及拓扑关系。这些多源异构的网络安全数据可以采用具有良好扩展性的通用 XML 格式进行表示，也可以采用如 STIX [2]的威胁情报标准格式，或通过 OWL 本体描述语言等可自定义的方式来描述。网络安全数据融合就是对上述各类信息源产生的数据进行过滤、相关和集成，从而形成针对网络安全事件、态势和知识的表示架构，这种架构适合于获得有关网络安全的解释和决策，达到系统的网络安全保护目标(如提高报警率或跟踪动态目标)，实现可靠的网络管理和安全控制等。

从发展历史来看，数据融合所采用的具体方法可分为三个阶段：基于统计学的融合、基于数据挖掘的融合、基于人工智能的融合，其融合效果不断提升，主要表现在隐含关系的发掘能力、网络安全活动

场景的拼接还原能力，以及对于数据内容的自主适应能力方面。下面我们分别针对这三个阶段，介绍数据融合所采用的关键技术方法。

## 3. 基于统计学的融合技术

统计学是经典的数学分支，在诸多领域有着广泛应用，对于需要进行不同维度的规律发掘和行为分类的场景尤为适合，本节介绍以下两类典型的基于统计学的融合方法，以及对于网络安全领域的适用性。

### 3.1. 假设检验技术

假设检验技术[3]是以统计学原理为基础，选择某种最优化假设检验(如卡方检验、最大似然分布检验等)的判决准则，通过执行多传感器数据的假设检验处理，从而获取综合的数据结论。

假设检验是数理统计学中根据一定假设条件由样本推断总体的一种方法。事先对总体参数或分布形式作出某种假设，然后利用样本信息来判断原假设是否成立，采用逻辑上的反证法，并依据统计学中的小概率原理。假设检验包括方差检验、t 检验、卡方检验等具体方法。通过清除落在假设检验中置信区间之外的数据点，进行数据异常值的检测、判断与清洗，从而逐渐推出符合置信要求的数据融合结果。

### 3.2. 滤波清洗技术

滤波清洗技术是将滤波过滤等相关技术由单一传感器扩展到多个传感器组成的数据信息网，用联合滤波相关算法执行多传感器的数据滤波，跟踪数据处理过程，从而获得最佳的数据清洗融合效果。

常用的滤波算法有卡尔曼滤波[4]、中值滤波[5]、粒子滤波[6]等，可以对各类数据进行平滑处理操作，这些方法在工业生产、自动控制、轨迹规划等方面有着很多用途，在网络安全领域同样有广泛的应用前景。例如针对突发的网络安全告警信息，利用上述算法进行平滑滤波，有助于清除部分攻击误报。这些误报有可能是由于检测规则准确性不足导致的，也有可能是缺乏数据背景，或检测算法对数据特性的适应性不足而引起的。

卡尔曼滤波的本质是参数化的贝叶斯模型，通过对下一时刻系统的初步状态估计(即状态的先验估计)以及测量得出的反馈相结合，最终得到该时刻较为准确的的状态估计(即状态的后验估计)，其核心思想即为预测+测量反馈，两者通过一个变化的权值相联系，使得最后的状态后验估计无限逼近系统准确的状态真值，这个权值即为卡尔曼增益。因此，卡尔曼滤波是典型的时序相关性滤波算法，这与网络安全领域中攻击活动的强时序相关性是吻合的。攻击者为了达到最终的攻击目的，通常会按照一定的先后次序依次执行相关操作，以逐步获取目标系统的各类信息和访问权限，因此，将卡尔曼滤波等算法用于网络安全数据的滤波清洗，将有助于发掘出各个攻击环节之间的时序相关性，从而支撑攻击链的还原和攻击场景重构。

上述两种数据融合技术均可应用于网络安全数据的清洗，剔除干扰数据，在判定数据准确性方面具有各自独特的优势。后者除用于数据清洗之外，特定的滤波清晰模型有助于构建多条数据之间的关联，实现时序数据的融合。

## 4. 基于数据挖掘的融合技术

数据挖掘尝试从海量数据中通过机器学习，获取数据中包含的知识。因此，数据挖掘本质上是知识挖掘。经典的数据挖掘算法有分类、关联、聚类等，在当今大数据时代背景下，数据挖掘方法已经广泛应用于各行各业，本节重点介绍其中基于聚类分析的数据融合方法。

聚类分析技术是以统计聚类分析或模糊聚类分析为基础，在多目标、多传感器大量观测数据样本的情况下，实现来自同一目标的数据样本自然聚类、来自不同目标的数据样本自然隔离，从而完成多目标

数据的融合。

聚类[7] [8]是按照某个特定标准(如距离准则)，把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象之间的相似性尽可能大，同时不在同一个簇中的数据对象之间的差异性也尽可能地大。即聚类后同一类的数据尽可能聚集到一起，不同类的数据尽量分离。在聚类分析中，常用的聚类方法有快速聚类(迭代聚类)和层次聚类。层次聚类容易受到极值的影响，并且计算复杂速度慢，不适合大样本聚类；快速聚类虽然速度快，但是其分类指标要求是定距变量。

k-means 是最为经典的聚类算法之一。由于该算法的效率高，所以在对大规模数据进行聚类时被广泛应用。目前，许多算法均围绕着该算法进行扩展和改进。k-means 算法以 k 为参数，把 n 个对象分成 k 个簇，使簇内具有较高的相似度，而簇间的相似度较低。k-means 算法的处理过程如下：首先，随机地选择 k 个对象，每个对象初始地代表了一个簇的平均值或中心；对剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇；然后重新计算每个簇的平均值。这个过程不断重复，直到误差函数收敛。

以 k-means 为代表的聚类算法，因其具备突出的无监督学习能力，能够在缺乏先验知识的前提下实现数据内涵的分割，因此在网络安全领域也开始得到了应用。在网络安全实际场景中，由于未知安全漏洞和新型攻击方式的不断出现，完全寄望于针对已知攻击行为的规则分类检测(典型的基于先验知识的检测技术)是不现实的，而无监督学习对于已知攻击和未知攻击一视同仁，而是从数据本身的特性来进行检测，摆脱了对于网络安全先验知识的依赖，从而获得了业界技术人员的广泛青睐。

## 5. 基于人工智能的融合技术

将人工智能技术应用于数据融合，对于解决融合中的不精确、不确定性信息有着很大优势，因此成为数据融合的重点发展方向。

神经网络(Neural Network)是通过机器学习实现人工智能的典型方法。它是一种由许多简单元组成的网络结构，这种网络结构类似于生物体的神经系统，用来模拟生物与自然环境之间的交互，借鉴生物体对于外界数据的分析判断能力，来实现数据的综合化处理。例如人体通过视觉、听觉、触觉、嗅觉、味觉等不同类型的多源数据，能够对外界事物产生综合画像，并作为指导自身作出反应的依据。神经网络是一个大的范畴，针对语音、文本、图像等不同的学习任务，衍生出了更适用于具体学习任务的神经网络模型，如卷积神经网络、循环神经网络等。

卷积神经网络(CNN)是一种前馈神经网络[9]，通常由一个或多个卷积层和全连接层组成，此外也会包括池化层。每个卷积层由若干卷积单元组成——可以想象成经典神经网络的神经元，只不过激活函数变成了卷积运算。卷积运算有其严格的数学定义，在 CNN 的应用中，卷积运算的形式是数学中卷积定义的一个特例，它的目的是提取输入的不同特征。CNN 结构相对简单，可以使用反向传播算法进行训练，这使它成为了一种颇具吸引力的深度学习网络模型。除了图像处理，CNN 也被应用到语音、文本处理等诸多其他领域。

循环神经网络(RNN) [10]也称为递归神经网络。RNN 和普通神经网络的数据单一方向传递不同。RNN 的神经元接受的输入除了“前辈”的输出，还有自身的状态信息，其状态信息在网络中循环传递。长短时记忆网络(LSTM) [11]可以被理解为是一种神经元更加复杂的循环神经网络，当处理的时间序列中间隔和延迟较长时，LSTM 通常比 RNN 效果更好。相较于构造简单的 RNN 神经元，LSTM 的神经元要复杂得多，每个神经元接受的输入除了当前时刻样本输入、上一个时刻的输出，还有一个元胞状态。LSTM 在很大程度上缓解了一个在 RNN 训练中非常突出的问题，即梯度消失/爆炸(Gradient Vanishing/Exploding)问题。这个问题不是 RNN 独有的，其他深度学习模型都有可能遇到，但是对于 RNN 而言特别严重。

无论是卷积神经网络，还是循环神经网络，包括长短时记忆网络，相关算法在网络安全领域都已经

开始得到应用。在利用上述算法进行数据融合之前，需要进行数据关联，以决定来自不同传感器的哪些网络安全数据属于同一目标。攻击者为达到一定的目标，会采取相应的策略，在不同的阶段实施不同的攻击行为，从而可能在不同的安全数据中留下攻击痕迹。这些痕迹对应的攻击事件应该不是孤立的，他们之间存在某种必然的联系。例如，具有一组相同源 IP 地址和攻击类型的攻击告警，很可能就是来自同一主机的探测攻击正在扫描同一子网上不同机器的相同端口；又如，类似 SYN Flood 之类的假冒地址攻击，即使不是来自同一源 IP 地址，但也可能从目标 IP 地址和攻击类型的关联中找到联系。因此，攻击行为体现在监测数据的互作用中，漏洞情报、攻击流量、攻击行为之间存在必然的关联性，可以完整的描述一个安全事件的起因、过程和结果。

类似这些网络安全数据项之间的联系，虽然通过专业安全人员的人工分析也可以获得，但由于网络安全数据量庞大，数据类型复杂，如果能够利用神经网络的机器学习能力实现自动化的提取，将对网络安全保护工作提供巨大的便利性。

## 6. 多源异构数据融合实践

基于对网络安全数据融合要素的梳理与切分，我们搭建了异构多源网络安全数据汇聚和清洗融合原型系统，在该系统中上述所提及各类算法均以融合插件的形式体现，该原型系统在某单位进行了应用示范，重点对 10 余类流量日志中产生的网络安全威胁攻击数据进行融合分析，其中基于统计学的数据融合技术在流量日志清洗过程中发挥了主要作用，特别是滤波清洗技术，对载荷上传行为数据是否威胁相关的判定上表现突出。基于数据挖掘的融合技术主要应用于由当前传统流量检测、日志管理设备所产生的多源异构日志的聚合，日志量依据融合维度、融合策略会出现较大幅度的压缩，百万级数据可聚合形成可由人工研判的十余条日志，大大降低网络安全威胁攻击数据的运维工作量。基于人工智能的技术主要用于攻击场景的构建、攻击行为信息的关联提取，通过该类技术的应用，可为网络威胁攻击的形式化表达、攻击线索拓展分析奠定基础。

## 7. 小结

数据融合是大数据治理中常见的处理过程，在社会各个领域都得到了广泛应用，采用的融合算法也经历了长期的发展过程，有着清晰的技术发展脉络。从技术原理而言，本文所介绍的数据挖掘技术本质上也是从统计学发展出来的，即便是当今发展最为迅猛的人工智能技术中仍然能够看到统计学的身影。

综上，本文提出了网络安全数据融合要素，阐述了统计学、数据挖掘、人工智能等一系列技术算法，并重点探讨了各类算法对于网络安全数据融合任务的适用性，并对网络安全数据融合技术在实际项目中的应用进行了简要描述。网络安全是新兴的大数据应用领域，随着相关技术的不断发展，网络安全数据的融合处理需求也日益高涨，数据融合技术模型必将在网络安全保障领域发挥越来越重要的作用。

## 基金项目

本论文得到“异构多源网安大数据采集汇聚和清洗融合技术研究”课题资助。

## 参考文献

[1] Lou, R.C. and Key, M.G. (1989) Multisensor Integration and Fusion in Intelligent System. *IEEE Transactions on Systems Man and Cybernetics*, **19**, 901-903. https://doi.org/10.1109/21.44007

[2] Barnum, S. (2020) STIX-Whitepaper. http://stixproject.github.io/getting-started/whitepaper/

[3] Rao, C.R. (1973) Linear Statistical Inference and Its Applications. John Wiley & Sons, Inc., New York. https://doi.org/10.1002/9780470316436

[4] Kalman, R.E. (1960) A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME, Jour-*

*nal of Basic Engineering*, **82**, 35-45. https://doi.org/10.1115/1.3662552

[5] Gonzalez, R.C. and Woods, R.E. (2002) Digital Image Processing. Addision-Wesley, Boston.

[6] Chan, V. (2013) Theory and Applications of Monte Carlo Simulations. Intech, Rijeka. https://doi.org/10.5772/45892

[7] Chen, M., Han, J. and Yu, P. (1996) Data Mining: An Overview from Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 866-883. https://doi.org/10.1109/69.553155

[8] Han, J.W. and Kamber, M. (2000) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

[9] Lecun, Y. and Bottou, L. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[10] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[11] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735