数据挖掘技术在网络安全中的应用

薛明霞

（中通服网盈科技有限公司 南京 210000）

摘 要 为助力互联网的稳定、可持续发展，提升各行业领域的网络安全等级，文中以“数据挖掘技术”在网络安全领域的应用为核心，探讨了其算法过程和风险识别能力。在此基础上，构建了一种网络运维管理系统，以期能为网络安全管理工作提供参考。

关键词： 数据挖掘技术；网络安全；聚类分析

中图分类号 TN915.08

Application of Data Mining Technology in Network Security

XUE Mingxia

（ZTO Service Netying Technology Co.，Ltd.，Nanjing 210000，China）

Abstract In order to help the stable and sustainable development of the Internet and improve the level of cyber security data in various industries, this paper focuses on the application of "data mining technology" in cyber security, and discusses its algorithm process, risk identification ability, etc. On this basis, a network operation and maintenance management system is constructed to provide reference for cyber security management.

Key words Data mining technology，Network security，Cluster analysis

0 引言

数据挖掘技术集成了人工智能、机器学习、统计学、数据库等技术，能在海量数据中发现有价值的信息数据，并深入分析与探寻网络中的潜在威胁[1]。为进一步发挥数据挖掘技术的应用价值，本文围绕数据挖掘技术在网络安全领域的实际应用需求，构建了一种网络安全运维管理系统，为网络故障问题的及时发现与处理提供了保障。

1 数据挖掘技术和网络安全分析

数据挖掘技术和网络安全分析领域的联系主要体现在3个方面。(1)应用场景。在实际应用中，数据挖掘与网络安全分析存在一定的重叠，如网络攻击行为的识别与分类、网络安全事件的预警及检测等。(2)算法及方法。数据挖掘及网络安全分析可应用相似的方法及算法，如支持向量机、神经网络、决策树等。(3)数据源。数据挖掘和网络安全分析都需要以大量的数据为输入，如网络日志、系统日志、应用日志等。

2 数据挖掘技术的核心算法

2.1 聚类分析

聚类分析是数据挖掘技术的重要组成，其应用原理为基于数据相似性的自动分组，即通过计算数据间的距离，将距离最近的数据划分为同一组别。聚类分析算法主要包含3个步骤。(1)数据预处理：转换原始数据为适合分析的格式；(2)距离计算：基于数据特征计算距离；(3)聚类计算[2]。

聚类分析的数据模型有两种。(1)凸包。凸包可以将数据集中的点围成一个凸多边形。凸包算法的核心思想是从数据集中选择一个点作为凸包的起点，然后逐个选择距起点最近的点，直到所有的点都被纳入凸包。(2)欧几里得距离。该方法需要计算两点间的距离，如式(1)所示：

$$d=\sqrt{\sum_{n}^{i=1}(x_i-y_i)^2} \tag{1}$$

其中，d代表距离；x_i、y_i代表两点。

2.2 决策树挖掘

决策树挖掘可以基于数据构建决策树。决策树可以将数据以树状结构表示出来，每个节点都代表不同的特征，分支代表不同特征的取值，叶片节点则代表一个决策。该算法主要包含4个步骤。(1)数据预处理：将数据转换为适合分析的格式。(2)特征选择：基于特征重要性选择最佳特征。(3)构建决策树：基于选择的特征递归地构建决策树。(4)决策树剪枝：为规避过拟合问题，需做好决策树剪枝[3]。决策树挖掘的数学模型有两种。

(1)信息增益。信息增益可用于代表一个特征能提供的信息量，如式(2)所示：

作者简介：薛明霞(1989—)，专科，助理工程师，研究方向为电子信息工程。

$$Gain(S)=I(D)-I(D|S) \quad (2)$$

其中，(S)代表一个特征；(D)代表数据集；$I(D)$代表数据集信息量；$I(D|S)$代表基于特征的数据集信息量。

（2）基尼指数。基尼指数是一种用于衡量特征纯度的指标。从数据集中随机抽取两个样本，对一个包含K个类别的数据集D而言，其基尼指数计算公式为：

$$Gain(D)=\sum_{k=1}^{K}Pk(1-Pk)=1-\sum_{k=1}^{K}p_k^2 \quad (3)$$

$$Gain(D,A)=\frac{|D_1|}{D}Gain(D_1)+\frac{|D_2|}{D}Gain(D_2) \quad (4)$$

其中，Pk代表类别k在数据集D中的比例。

基尼指数的取值范围一般在$[0,1]$之间，数值越小，代表数据集的纯度越高。

2.3 关联规则挖掘

关联规则挖掘主要用于发现数据之间的关联关系。其核心思想是通过统计数据的出现频率发现数据之间的关联关系。关联规则挖掘算法包含4个步骤。（1）数据预处理：转换原始数据为适合分析的格式。（2）频繁项集挖掘：通过*Apriori*算法找到频繁出现的项集。（3）关联规则生成：基于频繁项生成数据关联规则。（4）关联规则评估：通过信息增益、支出度对关联规则的有效性进行评估。关联规则挖掘的数学模型有以下两种。

（1）支持度。支持度是指一个项集在整个数据集中的出现频率，如式（5）所示：

$$Supp(X)=\frac{Count(X)}{Total} \quad (5)$$

其中，(X)代表一个项集；$Count(X)$代表项集(X)在数据集中出现的次数；$Total$代表数据集总体数量。

（2）信息增益。信息增益是指一个项集能提供的信息量，如式（6）所示：

$$Gain(X\rightarrow Y)=I(X)-I(X\cup Y) \quad (6)$$

其中，(X)代表一个项集；(Y)代表另一个项集；$I(X)$代表项集(X)的信息量；$I(X\cup Y)$代表项集$(X\cup Y)$的信息量。

3 数据挖掘技术在网络安全领域的应用

为解决人机交互、自动清理无用数据、网络故障自动挖掘等现实需求[4]，本文基于上述数据挖掘技术设计了一种网络安全运维管理系统。

3.1 结构设计

结合实际需求，本文基于B/S架构来设计网络安全运维管理系统。整个结构分为3个部分，分别为表示层、服务层以及数据层，如图1所示。

（1）表示层是人机交互界面，其程序编码基于ASP技术实现。用户可在该结构层面实现对系统的操作并展示运维管理信息。

（2）服务层是整个系统的核心，其主要作用是开发网络安全运维所需的功能，并基于这些功能进行程序编写。在服务层中，所有功能的开发需基于.NET平台展开，利用SQL Server Data Tools for VS智能开发工具，在数据库中实现源数据抽取，而后应用SQL Server Analysis Servers分析服务器的决策树算法、关联算法等进行数据挖掘、分析，构建所需的数据挖掘模型。随后，以该模型为基础，依托DMX指令检索模型。查询业务的封装则需借助C#语言完成，为网络安全故障分析预测模型的建设提供保证，完成基于Silverlight的应用功能开发。

（3）数据层。数据层由两部分构成。1）网络安全运维管理系统数据库主要为MySQL5.7数据库（一种开源性关系型数据库）。2）网络安全运维故障信息数据仓库主要用于挖掘网络安全故障信息，同时防止数据泄露或丢失[5]。

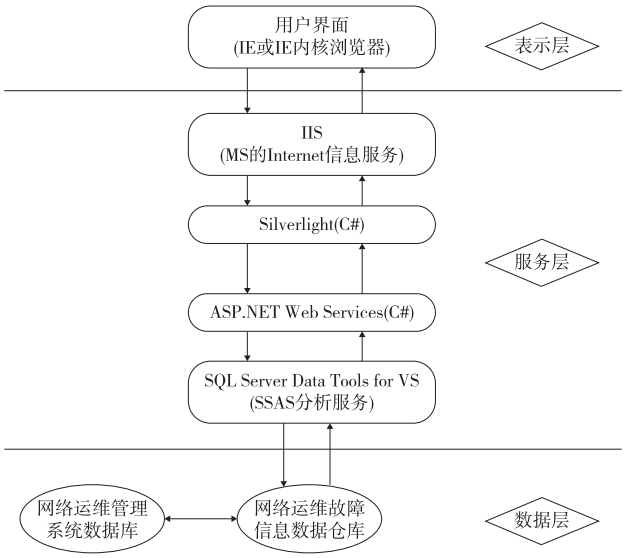

图1 网络安全运维管理系统的整体结构

3.2 功能规划

结合网络安全运维管理系统的功能需求，整个系统需划分为5个模块，如图2所示。

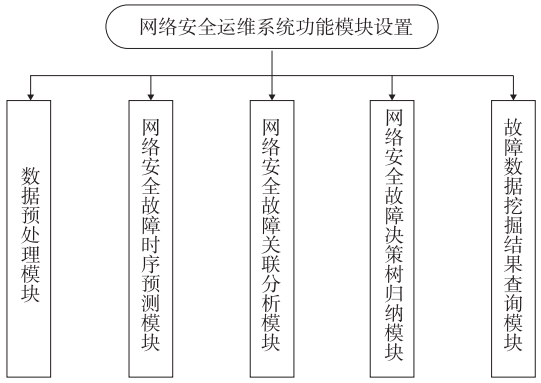

图2 网络安全运维系统功能规划

（1）数据预处理。数据预处理模块可以集中处理原始数据集，去除其中的噪声数据、失真数据，并修补缺失数据，提升数据集的合理性。在确保数据挖掘结果精准性的基础上，需注重缩短挖掘时间，提高系统运行效率。

（2）网络安全故障时序预测模块。在该功能模块中，首先需以网络安全故障类型、所处区域为基准，基于预设时间自动记录、统计故障数据及相关信息，而后生成时间序列。将故障数据上传至网络运维故障信息数据仓库后，故障数据的挖掘与分析即可通过时序算法实现，以便作出合理的判断，保证网络安全故障预测功能的正常运行。

（3）网络安全故障关联分析模块。在该功能模块中，需以数据仓库中的用户数据、网络故障数据为核心，通过一定的关联规则算法进行计算，并从不同的角度出发，对不同数据间的不同属性进行关联规则分析，确定潜在的故障关联，实现故障预测功能。

（4）网络安全故障决策树归纳模块。该功能模块的设计需基于数据仓库中的网络安全故障数据，基于决策树算法来分析故障数据的所有属性，并探寻同类型网络安全故障的共性、规律，为网络安全故障诊断提供支持。

（5）故障数据挖掘结果检索与查询模块。该模块可以为用户提供数据检索、查询等服务。用户在输入故障数据关键词后，即可实现检索结果的自动关联，所得查询结果也能通过图像、表格等形式完整呈现出来[6]。

3.3 数据库设计

数据库的开发应用了“前端Vue”“后端Java程序语言开发”“MySQL5.7数据库”“SSM框架”等技术。

（1）前端Vue应用是一种轻量级的JavaScript框架，能快速构建交互式的用户界面。同时，Vue还能提供便于使用的API，能使网络安全故障维护系统的设计者、开发者迅速创建可复用、组件化的代码。

（2）后端Java程序语言是一种跨平台编程语言，其自带工具生态系统和丰富的库，在企业级应用的开发中得到了广泛应用。Java在后端服务开发方面具备强大的可伸缩性和性能，能实现与其他多种技术栈的集成。在网络安全故障维护系统的设计过程中，可在加载阶段将类的class文件读入内存，并为之创建一个java.lang.Class对象（该过程由类加载器完成）。JVM在该阶段的主要作用是将字节码从不同的数据源（可能是class文件、jar包甚至网络文件）转换为二进制字节流并将其加载到内存中，生成一个代表该类的java.lang.Class对象。而后，进入连接阶段（可分为验证、准备、解析等阶段）。该阶段会将类的二进制数据合并到JRE中，以确保Class文件的字节流中包含的信息符合当前虚拟机的要求，且不会危害虚拟机自身的安全性。

（3）MySQL5.7数据库是一种开源性关系型数据库，在网络安全故障维护系统的设计中，数据库中的数据主要以结构化的形式存在，在相关领域中得到了广泛的应用，提升了数据库的检索速度与灵活性。

（4）SSM框架。在设计网络安全故障维护系统的过程中，主要涉及3个框架。1）Spring MVC框架，主要负责职责解耦；用户可以请求数据，并将请求地址发送至特定的服务端；系统接收请求后进行数据分析，之后返回数据，由页面视图渲染并展现经过处理的数据。2）Spring框架由7个模块构成，包括Spring Core、Spring Context、Spring DAO、Spring ORM、Spring Web、Spring AOP以及Spring Web MVC。3）Mybatis框架。该框架可对系统程序中的持久层、访问层进行抽象，并适当减少代码量，对数据库设计十分有益。

4 结语

数据挖掘技术在计算机网络安全维护领域有着至关重要的作用。本文结合网络安全故障维护的实际需求，设计了一种网络安全故障维护系统，以期能为该领域的发展提供助力。

参考文献

[1] 李娜.数据挖掘技术对烟草行业网络安全态势的评估方法设计研究[J].经济师，2024(5)：288-289.

[2] 王志刚，王辰阳，刘成.基于数据挖掘技术的网络运维管理系统研究[J].中文科技期刊数据库（全文版）工程技术，2024(6)：200-203.

[3] 钱祖良，朱铁良.大数据技术在计算机网络信息安全问题中的应用——评《计算机网络信息安全》[J].应用化工，2024，53(2)：511.

[4] 蒋科.基于数据挖掘算法的医疗网络安全风险评估[J].自动化技术与应用，2023，42(5)：99-102，127.

[5] 杨红艳.基于数据挖掘的能源互联网数据安全风险检测方法[J].信息技术与信息化，2023(7)：145-148.

[6] 赵相楠.基于数据挖掘的网络通信数据安全风险识别算法研究[J].计算机应用文摘，2022，38(15)：101-103.