（3）获取根节点交换机对应的 Bridge 信息，若 dot1dTpFdbPort 中无重复，则记录其对应端口的 MAC 地址 dot1dTpFdbAdress。若 dot1dTpFdbPort 中出现重复，则根据 ArpInfo_List 中端口索引、交换机直连关系 Switches_Lines、以及 dot1dTpFdbPort 最小重复次数为依据，确定根节点直连交换机 MAC 和 IP。记录交换机与交换机端口连线关系，并将新发现的交换机作为新的根节点交换机，递归执行（1）（2）（3）步骤；

（4）根据 ArpInfo_List 集合、ChildNet_List 集合，综合分析与之相连的子网或主机设备；

（5）遍历 ARP 表中未连接的离散点，获取其 ArpInfo_List 集合中的 IP、端口索引进行匹配，若其交换机 IP、端口索引完全匹配，则记录其连线关系；

（6）采用 ICMP、OSPF 等协议，验证拓扑分析的准确性，并加以修正。同时，确定设备节点的实时状态。

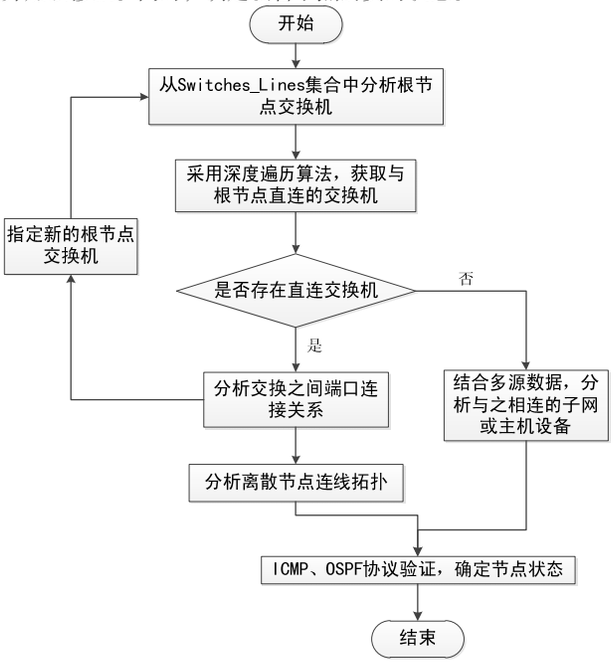

图 4 综合分析拓扑结构

3 结果与分析

本实验环境主要由 9 台交换机组成，一个核心交换机，多台接入交换机，在接入交换机上配置网关，连接各子网。

采用传统拓扑发现算法，其采集拓扑关系如下（图 5）。

实验分析发现，与核心交换机直连的还有 192.168.9.*的子网，且存在部分交换机发现不全等现象。

采用本文改进拓扑算法分析，其最终发现的拓扑图如下（图 6）。

实验结果与实际真实拓扑基本一致。经对比说明，改进拓扑发现算法能够更准确、全面发现拓扑结构。能够对全网拓扑发现起到支撑作用。

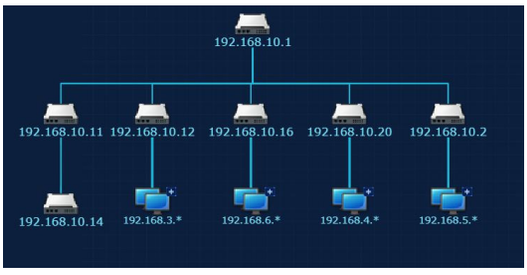

图 5 传统算法拓扑发现效果图

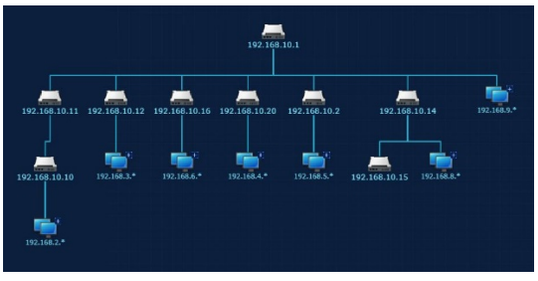

图 6 改进算法拓扑发现效果图

4 结束语

本文给出了一种基于多源拓扑自动发现算法。能够基于现有 ARP、SNMP 等协议的局限性，尽可能完整发现全局网络拓扑结构，能够有效过滤网关 IP，提供更为高效的拓扑发现和分析手段。在今后工作中，要尝试对含有防火墙、安全策略的网络拓扑结构进行研究，相信会有更多的优秀成果。

参考文献：

[1]张春强.深入理解 Net-SNMP[M].北京：北京机械工业出版社，2015.

[2]薛健.IP 级网络拓扑发现技术的研究与实现[D].哈尔滨：哈尔滨工业大学，2013.

[3]Nur A Y，Mehmet E T.Cross-AS（X-AS） Internet Topology Mapping[Z].Computer Networks，2018：132.

[4]Motamedi R，Rejaie R，Walter W.A.Survey of Techniques for Internet topo-logy discovery[J].IEEE Communications Surveys & Tutorials，2015，17（02：1044-1065.

[5]Kardes H，Gunes M，Oz T.Cheleby：A Subnet-Level Internet Topology Mapping System[C].Communication Systems and Networks（COMSNETS），2012：1-10.

[6]周长建.融合多协议的网络层拓扑发现算法研究[D].哈尔滨：哈尔滨工业大学，2017.

一种基于隐马尔可夫模型的口令猜测方法

◆王 蕊[1,2] 徐岳皓[1] 石 珺[1,2] 吕 博[1,2] 周 阳[1,2]

（1.中国电子科技集团公司电子科学研究院 北京 100041）

（2.中电科网络空间安全研究院有限公司 北京 100041）

摘要：随着信息时代的发展，信息安全尤其是口令的安全性研究成为热点问题。本文提出一种基于隐马尔可夫模型的口令猜测方法，通过为口令猜测问题建立隐马尔可夫模型，以大量的用户口令数据集为输入，以训练得到的初始状态概率、转换概率、观测概率等信息为依据，对数据集的口令字符组合、顺序等模式进行挖掘，最终以观测序列生成的方式产生新的预测密码，实现口令猜测。
关键词：口令猜测；隐马尔可夫；转移矩阵

信息安全问题受到越来越多的重视，用户认证机制成为保护用户隐私安全的一种常用方法[1~3]。其中，文本口令具有实现简单方便等的特点，是最为普遍使用的一种用户认证方式，然而却存在口令强度不够高的问题。因此，口令的安全性研究成为热点问题，其中利用不同方法实现对口令集的猜测[4-6]是一项重点研究方向，具有重要研究意义。一方面，从破解者的角度猜测用户口令可以使人们对口令的安全性有更深的认识；另一方面，猜测口令也可用于检测用户口令的强度，可方便用户在选择口令时避开容易被猜测到的口令，或在猜测口令结果的基础上更改口令以提升口令强度。

为了便于记忆，用户的口令一般不是随机的字符串，而是会包含一些特殊含义字符，例如，在选择口令时加入姓名缩写，生日等个人信息。传统研究[7, 8]在实现用户口令模式概率模型的训练时，需要利用用户的个人信息，但网上泄露的口令数据集往往并不包含用户信息。因此，为了提高模型的训练效率与预测精度，往往需要进行大量的前期调研，寻找与给用户口令匹配的信息。而且这种口令猜测方法更针对的是某个特定的用户，在猜测口令时，也要预先知道用户信息。综上，一般而言，实现口令猜测时只考虑了数字口令，但在当今时代，使用纯数字作为口令的人少之又少，现实意义不强。

尽管每个用户选择口令的偏好不同，但大量的用户口令数据集往往存在一定的统计特征和规律。因此，对大量用户口令数据集进行训练，可以挖掘出一些可预测的模式。本文利用隐马尔可夫模型进行口令猜测，基于口令字符之间的初始状态概率、转换概率、观测概率等信息对数据集的口令字符组合、顺序等模式进行挖掘可发现一定的规律。如此可见，根据隐马尔可夫模型猜测用户口令具有很强的现实意义。

1 基于隐马尔可夫模型的口令猜测方法

2.1 算法思想

本文提出一种基于隐马尔可夫模型的口令猜测方法，通过为口令猜测问题建立隐马尔可夫模型，以训练得到的初始状态概率、转换概率、观测概率等信息为依据，对数据集的口令字符组合、顺序等模式进行挖掘，最终以观测序列生成的方式产生新的预测密码，实现口令猜测。本文的优势在于不依赖如用户个人信息等外在条件，只根据泄露数据集中的所有用户口令来训练模型、猜测口令。本文的另一项优势在于不只针对数字口令，而是考虑了口令可能出现的所有字符类别，包括字母、数字、汉字、特殊字符这四种类别，可以更好地猜测复杂口令。

2.2 算法过程

本文主要通过构建隐马尔可夫模型，以{字母，数字，汉字，特殊字符}四种状态为模型对应的隐状态，以口令中出现的不同字符作为不同隐状态下的观测值，通过模型的训练生成不同观测序列，生成的观测序列即为猜测的口令。本文以口令数据集为训练数据，计算隐马尔可夫模型的转移概率、观测概率、初始状态概率，根据求得的初始状态概率选择初始状态，基于这个状态通过观测概率得到一个观测值；利用这个状态可以通过转移概率得到下一个状态，然后又可以继续得到下一个观测值；以此类推，直到把观测序列生成完毕。

算法流程图如图 1 所示，本文提出的基于隐马尔可夫模型的用户口令猜测方法包括下列步骤：

（1）数据预处理，实现输入数据的预处理，提取泄露口令数据集的所有密码，并以 0.75：0.25 的比例随机拆分为训练集和测试集；

（2）使用训练集数据构造隐马尔可夫模型$\lambda = [A,B,\pi]$=[转移概率，观测概率，初始状态概率]。

①设系统所有可能的状态集合为 $S = \{s_1,\cdots\cdots, s_4\}$ = {字母，数字，汉字，特殊字符}，即 4 种隐状态；

②令状态序列$E_i = (e_1,\cdots\cdots,e_{n_i})$代表训练集中第 i 个密码的字符状态序列，n_i为该口令的字符个数，再将所有状态序列拼接得到拼接的状态序列 $C = (c_1,\cdots,c_t,\cdots,c_T)$ ，其中c_t代表时刻 t 的状态，T 为训练集密码字符拼接序列总个数。

③观测序列 $O = (o_1, \cdots o_t\cdots, o_T)$，其中$o_t$为训练集密码序列中的对应字符，也为$c_t$的对应时刻 t 的显状态；

④计算可观测对象 $V = \{v_1,\cdots\cdots, v_q\}$ =口令拼接序列中的不同字符，v_i表示泄露口令数据集中出现的所有可能字符，q 为不同字符总个数，也为显状态个数。

⑤计算状态转移矩阵 A：

$A = [a_{ij}]$，$a_{ij} = P(e_{t+1} = s_j | e_t = s_i), 1 \le i,j \le 4$

其中a_{ij}表示任意时刻 t 的状态若为s_i，则下一时刻状态为s_j的概率，即任意时刻两种状态的转移概率。因为有 4 种隐状态，所以 A 是 4 行 4 列，为了避免计算到拼接时口令之间的状态转移，分别对每个状态序列E_i计算，最后累加状态转移次数并计算概率。

⑥计算观测概率矩阵 B：

$B = [b_{ij}]$，$b_{ij} = P(o_t = v_j | c_t = s_i), 1 \le i \le 4$，$1 \le j \le q$

其中 B 是 4 行 4 列，b_{ij}表示在任意时刻 t，若隐状态为s_i，则生成观察状态v_j的概率。

⑦计算初始状态概率为$\pi = (\pi_1, \cdots\cdots, \pi_4)$，用于表示初始时刻各状态出现的概率（口令首字符的状态概率），其中$\pi_i = P(e_1 = s_i), 1 \le i \le 4$=即 t=1 时刻状态为s_i的概率。

（3）对用户口令的长度进行统计，计算不同长度的口令使用频次，以此获得用户口令长度概率分布 P(len)，表示长度为 len 的密码的使用概率。

（4）按照以下过程生成预测密码：

①依概率分布 P(len)生成可能的密码长度 len；

②根据初始状态概率π生成初始状态 S；

③基于该初始状态 S，通过观测概率 B 生成观测值 O；

④通过转移概率 A 计算下一个状态；

⑤重复上述步骤（3-4），可以继续得到下一个观测值，直至观测值长度累计为 len 为止。

（5）重复步骤 4，直至生成 m 个猜测口令，构成猜测口令集。其中 m 为测试集口令数目；

（6）对比猜测口令集和测试集，计算猜测口令集的命中率，从而得出猜测正确率。

2 优势分析

本文基于隐马尔可夫模型生成猜测口令，通过对口令数据集的挖掘学习得到隐马尔可夫模型，以训练得到的初始状态概率、转换概率、观测概率等信息为依据，结合观测序列生成的

方式产生新口令，从而实现口令猜测。本文可实现四种类型（字母、数字、汉字、特殊字符）字符组合的口令猜测，且该方法不局限于某个具体用户，而是猜测用户们可能使用的口令，并且随着数据集的增加，模型精度会越来越高，猜测的口令也会更符合人类习惯。

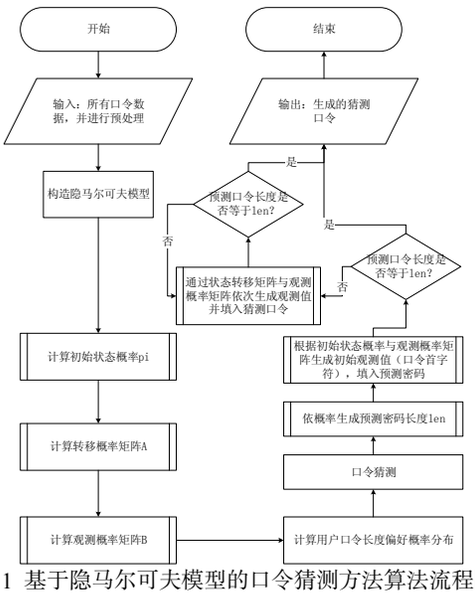

图 1 基于隐马尔可夫模型的口令猜测方法算法流程图

3 结束语

与现有技术相比，本文提出的技术方案中的模型建立只依赖泄露口令数据集内所有用户口令，避免了对用户信息的调研与整理，减少了人工工作量。本文将口令猜测问题转化为基于隐马尔可夫模型的观测序列生成问题，基于模型的初始状态概率、状态转移概率矩阵、观测概率矩阵，按照概率的形式生成猜测口令，可支持字母、数字、汉字、特殊字符等四种字符组合口令的猜测。此外，通过计算目标用户口令在隐马尔可夫模型下出现的似然度，可为检测目标用户口令的安全强度提供了一个新思路。

参考文献：

[1].网络信息安全技术综述[J].甘肃科技，2009，25（17）：29-33.

[2]王平，汪定，黄欣沂.口令安全研究进展[J].计算机研究与发展，2016，53（10）：2173-2188.

[3]刘功申，邱卫东，孟魁，等.基于真实数据挖掘的口令脆弱性评估及恢复[J].计算机学报，2016，39（3）：454-467.

[4]周浩，王靖康，王博，等.明文口令生成模型研究综述[J].计算机工程与应用，2018，54（4）：9-16.

[5]Yan J J，Blackwell A F，Anderson R J，et al. Password Memorability and Security：Empirical Results[J].IEEE Security and Privacy Magazine，2004，2（5）：25-31.

[6]Bonneau J，Herley C，Van Oorschot P C，et al. Passwords and the evolution of imperfect authentication[J]. Communications of the ACM，2015，58（7）：78-87.

[7]滕南君，鲁华祥，金敏，等.PG-RNN：一种基于递归神经网络的密码猜测模型[J].智能系统学报，2018，13（06）：29-36.

[8]周环，刘奇旭，崔翔，等.基于神经网络的定向口令猜测研究[J].信息安全学报，2018，3（05）：29-41.

一种设备状态监测的贝叶斯正则化 BP 神经网络

◆孙发友　蒙祖强

（广西大学计算机与电子信息学院　广西　530004）

摘要：设备全过程监测是有效消除设备隐患的重要环节，依据设备运行数据快速、准确、可靠的创建运行状态模型是建立故障早期预警的关键。以火电厂为例，由于火电厂现场生产环境复杂，正常状态设备会产生异常数据，本文首先将采用两倍差法进行数据清洗剔除异常数据，其次利用因子分析算法提取综合指标，最后基于贝叶斯正则化 BP 神经网络建立预测状态模型。将模型输出预测值与实测值进行计算产生设备状态信息。实验结果表明，预测模型输出值与实际状态值匹配度超过 90%，已经达到实际生产环境要求，具有推广价值。

关键词：数据清洗；综合指标；贝叶斯正则化；BP 神经网络；预测模型

现场运行设备的状态是一个实时动态变化的过程，通过在整个电厂建立主要生产设备运行状态的早期故障监测预警，构建起厂级的数据挖掘和设备健康管理系统，从而大大降低设备潜在的事故发生概率，缩短关键设备的非计划停机时间，提高设备的可靠性和可利用率[1]，降低设备维护运行成本。为此，需要解决一下问题：

（1）设计从设备日常运行的海量数据中自动过滤异常数据的算法。

（2）每个设备实时监测了多方面的特征参数，基于筛选过的数据，设计提取代表设备状态的综合指标的算法，从而优化监测数据处理的维度和监测成本。

（3）建立设备状态预测模型，能依据现有的设备状态值输出预测设备状态值。利用预测值与实际值的相似度产生设备状态信息。

设备状态数据为结构化数据，并且其数据分布为非正态分布。为了达到实验目标，采用基于贝叶斯正则化 BP 神经网络模型，加快数据收敛，同时该状态预测模型需要能够进行数据的自我学习，即依据自身已有的数据预测未来的数据。实验表明该模型具