海 南 大 学

毕 业 论 文（设计）

题　　目：基于 k-means++的聚类算法研究

学　　号：20151681310378

姓　　名：杜欣然

年　　级：2015 级

学　　院：计算机与网络空间安全学院

系　　别：计算机系

专　　业：计算机科学与技术

指导教师：杨厚群

完成日期：2019　年　5　月　5　日

摘　要

随着计算机应用日趋广泛和深入，网络安全问题也更加复杂和突出，它不仅仅关系到我们个人的隐私，更关系到商业利益乃至国家安全。现有的各种安全技术可以保证网路环境一定的安全性，但由于攻击手段层出不穷，现有安全技术也无法保证绝对安全。如何监控网络攻击并做出相应的防御已成为当今网络安全需要解决的重要问题。数据是网络时代的产物，传统的基于数据挖掘的入侵检测模型完全依赖于数据挖掘算法对已标记数据集中数据样本的学习。数据样本标记的正确性和纯度对于构建有效的入侵检测系统至关重要。但网络中的数据量巨大，想要得到纯净的样本数据代价极大。因此寻找一种对数据集要求不那么高的入侵检测方法至关重要。

基于以上背景，本文基于聚类技术进行入侵检测研究。本实验是在Windows平台，采用Anaconda集成环境，对KDD99数据集进行处理分析，在未知数据样本类别的情况下，通过计算样本彼此间的距离来估计样本所属类别，最终得出聚类结果。

[关键词]： 数据；网络安全；聚类

Abstract

With the increasing and deepening of computer applications, network security issues are more complex and prominent. It is not only related to our personal privacy, but also to commercial interests and even national security. The existing security technologies can guarantee a certain security of the network environment, but because of the endless stream of attacks, existing security technologies cannot guarantee absolute security. how to monitor network attacks and make corresponding defenses has become an important issue that needs to be solved in today's network security.

Data is a product of the network age. The traditional data mining-based intrusion detection model relies entirely on data mining algorithms for learning data samples in tagged data sets. The correctness and purity of the data sample markers is critical to building an effective intrusion detection system. But the amount of data in the network is huge, and it takes a lot of money to get pure sample data. Therefore, it is important to find an intrusion detection method that requires less high data sets.

Based on the above background, this paper conducts intrusion detection research based on clustering technology. This experiment is based on the Windows platform, using the Anaconda integrated environment to process and analyze the KDD99 dataset. In the case of unknown data sample categories, the sample is classified by calculating the distance between the samples, and finally the clustering result is obtained.

[Key Words]: data;network security;clustering

目　　录

1 绪论

1.1 选题的背景及意义

1.1.1 选题的背景

由于计算机技术的不断更新，人们已经完全进入了互联网时代。与此同时，网络安全问题日益严重。互联网的广泛开放和移动支付的普及使得一些重要领域受到了越来越多的入侵攻击。网络安全不仅是一个技术问题，而且已成为全球主要的信息安全问题。

最大的勒索病毒比特币勒索软件在2018年袭击了世界，造成了无法估量的损失。尽管网络安全问题得到广泛的关注，但此类事件并未减少。因此，构建数据分析模型使安全人员能够及时检测入侵十分重要，性能好的数据分析模型不仅预测准确，还能节省预测的时间成本。

1.1.2 选题的意义

目前世界上入对侵检测的研究涉及众多学科，如统计学、数据挖掘、机器学习等。为了获得更好的入侵特征，本课题基于聚类分析，从网络信息安全领域的先验知识入手，提取那些反映出网络异常行为的特征，然后使用恰当的算法对处理后的数据进行挖掘。

本课题的研究重点是基于网络的无监督异常检测系统的数据分析方法。由于入侵检测要分析的数据量巨大，数据特征复杂、维度较高，因此在聚类分析前要对数据进行大量处理，以便于观察和分析聚类结果。

1.2 国内外发展状况及其研究方向

入侵检测系统，简称IDS（Intrusion Detection System），是网络空间安全中的一个重要问题。它是一个实时监控网络或网络内部的系统。一旦发现攻击尝试或攻击，IDS将发出警告并提示以确保网络安全。由于传统防火墙大多使用静态防御并且缺乏实时警告，因此它们也无法攻击深度攻击。而 IDS可以实时响应入侵。

现如今，国外的一些研究机构对入侵检测的研究水平较高，普渡大学、加州

大学的Davis分校等在此领域处于国际领先高度。国外的一些知名厂商如Cisco等对于此的研究也很深入。对于IDS的研究国内起步的较晚，但发展迅速，许多国内厂商已经转向入侵检测领域，并且还推出了自己的网络安全产品，如中科网络的“天眼”入侵检测系统，启明星辰的SkyBell和绿盟网络入侵检测。然而，由于当前入侵检测技术中的各种缺陷，并且各种类型的攻击不断更新，误报率和误报率都很高。因此，需要进一步提高入侵检测的准确性。就目前而言，模式匹配技术仍然是大多数成熟商家用作IDS的主要技术。

1.3 本课题的研究概况

1.3.1 研究内容

1.3.2 研究重点及难点

1.4 论文的结构及内容安排

第一章为绪论。介绍了课题的背景和本课题重要性，除此之外，基于国内外对网络入侵检测的研究现状，简要介绍了研究的内容、难点以及本课题的主要工作。

第二章为本课题使用的相关技术的概述。介绍了开发环境、开发语言以及配套技术，为本课题做准备工作。

第三章为数据集预处理。介绍了分析数据集的方法并通过特征工程执行特征提取和特征降维。

第四章为聚类算法设计。本章分析k-means算法并对其进行优化以实现更好的聚类结果。

第五章为总结。指出在数据预处理中忽视的细节和需要进行改进的地方，还有对本次实验的总结归纳。

2 相关技术概述

2.1 数据挖掘概述

2.2 聚类的基本知识

2.2.1 聚类方法描述

2.2.2 常用的聚类分析方法

2.3 相关工具及技术

2.3.1 Anaconda

2.3.2 Python sklearn

2.4 本章小结

3 KDD99数据集处理

3.1 KDD99数据集简介

KDD99数据集是模拟数据集，模拟美国空军局域网搜集的大概九周的网络连接数据，可分为两部分：带有标识的训练数据、未加标识的测试数据。为了检测数据模型，测试数据中包含了训练数据中没有的数据类型，以便更接近真实的入侵检测。本次实验采用KDD99的训练数据集。正常识别类型和22种训练攻击类型包含在训练数据集中，如图所示。此外，有14种攻击仅出现在测试数据集中而不在训练集中。

表 1 训练数据集标识类型1

标识类型	含义	具体分类标识
Normal	正常记录	Normal
DOS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	buffer_overflow、loadmodule、perl、rootkit

表 2 训练数据集标识类型2

标签	类别	训练集（10%）	测试集（Corrected）
0	NORMAL	97278	60593

数据特征：KDD99训练数据集中有42维特征，其中前41维特征是连接记录的固定特征，最后一维是类标识符，用于指示连接记录是正常还是特定攻击类型。在前41维特征中，9个特征属性是离散的数据，而其他属性是连续的数据。如下图3-2所示。

标识类型	含义	具体分类标识
Normal	正常记录	Normal
DOS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	buffer_overflow、loadmodule、perl、rootkit

图 3-1 训练数据集标识类型

2, tcp, smtp, SF, 1684, 363, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 104, 66, 0.63, 0.03, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, normal.

图 3-2 KDD99训练数据集

3.2 数据预处理

3.2.1 观察数据

3.2.2 离散型数据预处理

离散数据，也称字符型数据。计算机不能直接处理字符型数据，因此我们需要在算法开始之前对数据进行一系列的特征处理、转换。

针对KDD99数据集中的四类字符型数据做如下处理:

3.2.3 连续型数据预处理

(1) 设置变量X为数据集，变量Y为标签：

```
data = df.values
X = data[:, 0:39]
X
```

```
array([[0, 1, 22, ..., 0.0, 0.0, 0.0],
       [0, 1, 22, ..., 0.0, 0.0, 0.0],
       [0, 1, 22, ..., 0.0, 0.0, 0.0],
       ...,
       [0, 1, 22, ..., 0.01, 0.0, 0.0],
       [0, 1, 22, ..., 0.01, 0.0, 0.0],
       [0, 1, 22, ..., 0.01, 0.0, 0.0]], dtype=object)
```

图 3-3 数据集X

```
Y = data[:,39]
Y
```

```
array(['normal.', 'normal.', 'normal.', ..., 'normal.', 'normal.',
       'normal.'], dtype=object)
```

图 3-4 数据集Y

(2) 设置变量X为数据集，变量Y为标签：

进行数据标准化的原因：对于同一个特征来说，不同的样本中的取值有可能会相差很常大，一些异常的数据会误导模型的正确训练；除此之外，如果数据的分布很分散也会影响训练结果。以上两种数据的数据方差会非常大。此时，我们可以将特征中的值进行标准差标准化，即转换为均值为0，方差为1的正态分布。

其原理是

$$x^* = \frac{x - \overline{x}}{\sigma}$$

其中，x^*为原始数据的均值,σ为原始数据的标准差。σ反应了给定数据距离其均值标准差的大小，高于平均值的数据将获得正标准化分数，反之亦然将获得负标准化分数。

本实验使用python中的sklearn.preprocessing.StandardScaler类，通过 StandardScaler模块计算标准化。标准化数据如下：

```
# 数据标准化
from sklearn.preprocessing import StandardScaler
sScaler = StandardScaler()
rescaleX = sScaler.fit_transform(X)
rescaleX
```

```
array([[-0.06779179,  0.9257548 , -0.10406721, ..., -0.46320296,
        -0.25203979, -0.24946427],
       [-0.06779179,  0.9257548 , -0.10406721, ..., -0.46320296,
        -0.25203979, -0.24946427],
       [-0.06779179,  0.9257548 , -0.10406721, ..., -0.46320296,
        -0.25203979, -0.24946427],
       ...,
       [-0.06779179,  0.9257548 , -0.10406721, ..., -0.43695069,
        -0.25203979, -0.24946427],
       [-0.06779179,  0.9257548 , -0.10406721, ..., -0.43695069,
        -0.25203979, -0.24946427],
       [-0.06779179,  0.9257548 , -0.10406721, ..., -0.43695069,
        -0.25203979, -0.24946427]])
```

图 3-5 标准化后的数据

(3) 对数据进行归一化：

进行数据归一化的原因：拿到数据样本时，数据的单位往往不一致，在计算期间将数据缩放到特定间隔。处理用于比较和评估的指标时，可以将数据的单位限制去除并且将其转换为无量纲值，从而比较和加权不同单位或大小的指标。

3.2.4 特征降维

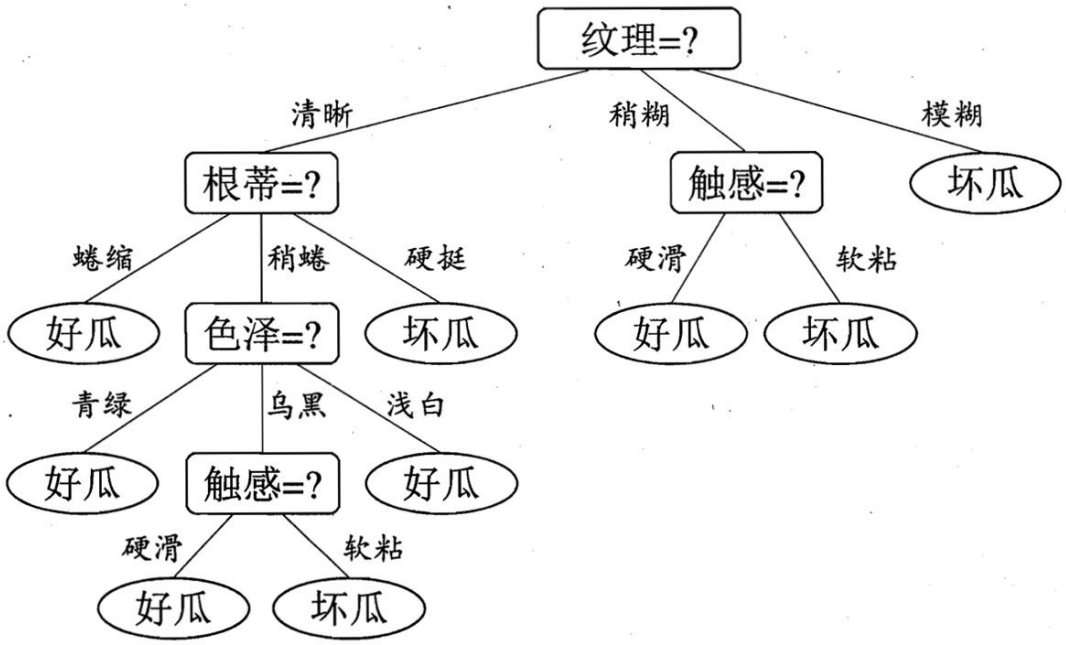

图 3-6 西瓜书

西瓜书 [1]中写道:

> 归纳（induction）和演绎（deduction）是科学推理的两大基本手段，归纳是从特殊到一般的“泛化”（generation）的过程

3.3 本章小结

总结一下。

[1]周志华，《机器学习》

4 基于聚类算法的入侵检测研究

4.1 k-means算法介绍

k-means算法是一种简单的划分聚类算法，它通过计算样本之间的距离大小将样本集划分为k个簇。聚类目标是使同一个簇中的相关对象尽可能相互“接近”，而不同簇中的对象尽可能地“远离”。

Algorithm 1: kmeans

Data: 样本数据集 D, 聚类簇数 k
Result: 聚类集合
$r \leftarrow t$;
$\Delta B^* \leftarrow -\infty$;
while $\Delta B \leq \Delta B^*$ *and* $r \leq T$ **do**
  $Q \leftarrow \arg\max_{Q \geq 0} \Delta B^Q_{t,r}(I_{t-1}, B_{t-1})$;
  $\Delta B \leftarrow \Delta B^Q_{t,r}(I_{t-1}, B_{t-1})/(r-t+1)$;
  if $\Delta B \geq \Delta B^*$ **then**
    $Q^* \leftarrow Q$;
    $\Delta B^* \leftarrow \Delta B$;
  end
  $r \leftarrow r+1$;
end

从Algorithm 1中可得 k-means 算法有以下缺点：

(1) 聚类中心的k数量需要提前给出，但实际运用中，对于给定数据确定k值十分困难。很多时候我们不知道应该将这些数据集划分成几类最佳。

(2) k-means算法的聚类中心是随机选择的，选择不好的初始聚类中心，可能导致完全不同的聚类结果。

4.2 最佳k值选择

5 总结

致　谢

谢谢杨老师

参考文献

[1] Zeng, Z., Wang, J.: Design and analysis of high-capacity associative memories based on a class of discrete-time recurrent neural networks. IEEE Trans. Syst. Man Cybern., Part B, Cybern. 38(6), 1525–1536 (2008)

[2] Zeng, Z., Wang, J.: Design and analysis of high-capacity associative memories based on a class of discrete-time recurrent neural networks. IEEE Trans. Syst. Man Cybern., Part B, Cybern. 38(6), 1525–1536 (2008)

附　录