

一种改进的聚类方法在异常检测中的应用

李 涵

(北京信息科技大学 理学院, 北京 100192)

摘 要: 入侵检测系统是一种积极主动的安全防护技术,它是信息安全保护体系结构中的一个重要组成部分.异常检测是入侵检测的一种方法,因其能够检测出未知的攻击而受到广泛的研究.以基于数据挖掘的异常检测技术为研究内容,以提高异常检测的检测率、降低误报率为目标,以聚类分析为主线,提出了一种改进的聚类检测算法和模型,并进行仿真实验.算法首先去除了数据集中明显的噪声和孤立点,通过分裂聚类、合并聚类以及利用超球体的密度半径确定 k 个初始聚类中心,以减小初始 k 值的选取对聚类结果造成的影响,提高异常检测效率,并以此构造入侵检测模型.利用 KDD CUP 1999 数据集对模型进行实验测试,并对改进算法的效果进行了对比和分析.实验证明,新的检测系统具有良好的性能.

关键词: 聚类;入侵检测;K-平均;异常检测

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2010)08-0066-04

Implementation of an Anomaly Detection Model Based on Improving Clustering Analysis

LI Han

(School of Science, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: IDS(Intrusion Detection system) is an active and driving defense technology. This paper mainly focuses on intrusion detection based on data mining. The aim is to improve the detection rate and decrease the false alarm rate, and the main research method is clustering analysis. The algorithm and model of ID are proposed and corresponding simulation experiments are presented. Firstly, a method to reduce the noise and isolated points on the data set was advanced. By dividing and merging clusters and using the density radius of super sphere, an algorithm to calculate the number of the Cluster Center was given. By the more accurate method of finding K clustering center, an anomaly detection model was presented to get better detection effect. This paper used KDD CUP 1999 data set to test the performance of the model. The results show the system has a higher detection rate and a lower false alarm rate, it achieves expectant aim.

Key words: clustering analysis; intrusion detection; K-means; anomaly detection

1 引言

入侵检测系统^[1]是一种积极主动的安全防护技术,它是信息安全保护体系结构中的一个重要组成部分.由于新的攻击方法的不断出现,尤其是一些互相协作的入侵行为的出现,给入侵检测领域的研究带来了新的课题.传统的入侵检测技术在扩展性

和适应性上已不能应付越来越复杂的攻击方式.近年来,数据挖掘(Data Mining)^[2]是引人关注的应用新技术之一,它主要融合了人工智能、统计学、数据库、机器学习等领域的方法与技术,利用分析工具从大量数据中提取隐含其中的,人们事先不知道的,但又是潜在的有用信息和知识.而入侵检测过程是利用所采集的大量数据信息,如主机系统日志、审计记

收稿日期: 2009-06-06; 修回日期: 2009-07-06

基金项目: 北京市教委科技创新平台基金项目(PXM2008_014224_067420),北京信息科技大学科研水平提高项目(5028123900)

录和网络数据包等,对其进行分析以发现入侵或异常的过程.从以数据为中心的观点看,入侵检测过程即是数据分析的过程.因此,在入侵检测系统中,可利用数据挖掘技术,对海量网络数据分析处理,从中提取尽可能多的隐藏的安全信息,抽象出有利于比较和判断入侵特征的计算模型.本文在数据挖掘和入侵检测理论的基础上,提出了一种基于聚类算法进行异常检测的入侵检测模型,对传统的 K-Means 算法进行改进,并以此构造入侵检测模型.利用 KDD CUP1999 数据集对模型进行实验测试,并对改进算法的效果进行了对比和分析.实验证明,新的检测系统具有良好的性能.

2 基于聚类的异常检测算法

聚类算法涉及很多领域,包括数据挖掘、统计、机器学习、空间数据库技术等.利用聚类分析方法进行基于异常的入侵检测不需要对数据对象作正常或异常标记,而是基于以下两点假设:(1)在整个系统及网络活动监测过程中,正常活动事件的数目要远远多于异常(入侵)活动事件的数目;(2)异常(入侵)事件与正常活动事件间存在着实质的差异.本文的基本思想就是由于入侵行为和正常的行为不同并且数目相对很少,因此对活动事件数据进行聚类分析,挖掘出其中表现异常的事件,并将其(如果存在的话)视为入侵活动事件.

K-Means^[3]是一种常用在异常检测中的聚类分析方法,它是 J. B. MacQueen 在 1967 年提出的一种基于划分的动态聚类算法.但是, K-Means 算法应用在异常检测中时却存在如下问题^[4]:

(1) 无法确定聚类个数 k 值

在利用 K-Means 算法解决问题的时候,很多时候并不知道给定的数据集应该分成多少个类最合适, k 值的大小作为 K-Means 算法的输入参数,对最终聚类结果有很大的影响,但是并没有公认的确定的 k 值的方法.

(2) 依赖初始聚类中心的选取

在 K-Means 算法中,初始聚类中心的选择对聚类结果有较大的影响,若初始化落在了一个局部极小点附近,就会造成算法在局部极小处收敛.

(3) 孤立点对聚类结果的影响

孤立点定义为在基于某些度量标准下,不符合数据的一般模型的一些数据对象.在入侵检测中,孤立点则意味着入侵. K-Means 算法对于数据集中孤立点很敏感,少量的这类数据将对聚类结果产生较

大的影响,孤立点的存在会使平均值得到很大的偏离,影响聚类准确度.

本文在对 K-means 算法做了大量研究的基础上,改进了对孤立点的处理、 k 值的选取和聚类中心的选择方法,将改进的 K-means 聚类算法应用到对入侵数据的检测中去,可以较好地解决传统入侵检测模型存在的缺陷.

2.1 对孤立点的处理

在系统和网络行为记录集中,正常行为记录数要远大于异常行为记录数,相对于异常行为记录来说,正常行为记录的分布应该更密集,孤立点被认为是具有入侵行为特征的记录.针对经典 K-Means 算法易受孤立点影响这一缺点对算法做了进一步改进,以减少噪声和孤立点对聚类效果的影响.对每个点 i ,找出其和其他点的距离和 D_i ,并计算出距离均和 H ,如果 D_i 大于 H ,则认为这个点是孤立点.其中 n 为样本数据 d 为数据的维数.具体定义如下:

$$D_i = \sum_{j=1}^n \sqrt{\sum_{k=1}^d (x_{ik} - x_{jk})^2} \quad (1)$$

$$H = \sum_{i=1}^n \frac{D_i}{n} \quad (2)$$

2.2 k 值的确定

K-Means 算法通过欧氏距离来划分类,假设合理的聚类数为 k ,若初始值选择 k_1 并且 $k_1 < k$,则说明至少有两个合理划分的类被归结成了一类.若初始值选择 k_2 并且 $k_2 > k$,则说明至少有一个合理划分的类被再次划分成了若干个类.在本文中取 $k_2 = \lfloor \sqrt{n} \rfloor$,作为 k 值的上限,其中, n 为样本集中样本的数目.通过文献[5]中的方法对聚类结果进行合并和分裂,以求得合适的 k 值.

2.3 选取适当的初始聚类中心

本文提出了一种寻找 k 个聚类中心的方法,从而可以得到更好的划分效果.将去除孤立点后得到的样本集 A' 作为聚类的初始样本集,计算 k 个聚类中心.

首先给出点的密度定义.对空间中任一点 p 以及距离 r ,以 p 为中心,以 r 为半径作超维球,落在该球内的数据点的个数称为点 p 关于距离 r 的密度,记作 $D(p, r)$.算法的基本思想是:给定一个距离阈值 R ,按照点的密度的定义,计算每一个数据对象关于距离阈值的密度,把对象按照密度排序,尽量选中密度较大的 k 个对象作为初始聚类中心.

初始聚类中心选择算法:

输入:有 N 条记录的数据集 D ,密度半径 R ;

输出:初始聚类中心集合 S

```
(1) 初始化初始聚类中心集合  $S$  为空
(2) for (读取数据集  $D$  中的每一个记录  $p_i$ )
    计算该记录的点密度;
(3) 按照点密度大小对集合  $D$  进行降序排序;
(4) while ( $D$  中仍有未读取记录) {
    读取当前记录;
    if (为第一和第二次读取的记录)
        将该记录加入到集合  $S$  中;
    else {
        计算当前记录与  $S$  中每个初始聚类中心的距离;
        把其中的最小距离保存在  $\text{dist-min}$  中;
        if ( $\text{dist-min} < S$  中记录间的最小距离)
            if ( $S$  中记录个数  $< k$ )
                把当前记录加入到  $S$  中;
            else if ( $S$  中记录个数  $= k$ )
                删除  $S$  中距离最大值点对中点密度较小的记录;
    }
}
```

其中,距离阈值 R 的取值方法如下:

```
(1) 对数据集进行标准化处理;
(2) 选择特征属性;
(3) 对正常数据和各种攻击类型数据各取 1 000 条记录保存在不同的表中,计算每一种类型数据的平均欧式距离并保存;
```

(4) 在混合类型数据的样本集中(包含正常数据和各种攻击数据的混合样本集),把平均欧式距离较小的类型数据的平均欧式距离作为 R 的初始取值,在此基础上实验得到最佳取值。

2.4 基于改进的 K-Means 的异常检测算法

结合以上研究,给出改进的 K-Means 算法流程如下:

```
(1) 扫描一次数据集  $A$ ,并计算出每个数据点的距离和  $D_i$  与距离均和  $H$ ;
(2) for(每个数据点  $i$ )
    如果  $D_i > H$  则视该点为孤立点;
(3) 去除  $A$  中的孤立点数据,得到新的数据集  $A'$ ,并记录  $A'$  中的样本个数  $M$ ,输出孤立点;
(4) 输入  $K$  的初始值  $K = M$ ,输入样本集合  $A'$ ;
(5) 运行 K-Means 得到  $K$  个聚类;
(6) 合并、分裂聚类;
(7) 输出  $k$  的值;
(8) for(读取样本集合  $A'$  中的每一个数据  $p_i$ )
    计算该数据的点密度;
(9) 按照点密度大小对集合  $A'$  进行降序排序;
(10) 扫描  $A'$  中的每一个数据,计算距离,获得  $k$  个初
```

始聚类中心;

(11) 聚类。

运用聚类算法的结果产生了若干个聚类,每个聚类中包含部分的连接记录。根据之前我们的假设,正常和异常的连接记录存在着实质性的差异,因此应该处于不同的聚类之中。又因为正常连接记录的数目远远大于异常连接记录的数目,我们可以设定一个阈值 N ,如果聚类中的连接记录(即数据对象)的个数大于这个阈值 N ,则把它标记为正常类,否则标记为异常类。

3 实验结果与分析

为了验证本文提出的基于聚类的异常入侵检测算法,选用 KDD Cup1999^[6] 实验数据。KDD CUP 1999 数据集中的入侵数据所占比例远大于正常情况下入侵记录比例。因此,需要把数据集中的很多入侵数据过滤掉,使得入侵数据占总数据量的 1% ~ 2%。

为了对算法进行正确而有效的评估,把数据挖掘算法的检测率和误检率两个方面的值作为算法性能的度量标准。检测率定义为算法检测到入侵记录数目占数据集中入侵记录总数的百分比;而误检率定义为被误认为是入侵的正常记录数占数据集中正常记录总数的百分比。

为了方便仿真实验,采用随机抽取的 5 组 10 000 个样本数据集,每个数据子集中正常数据记录个数占本子集所有数据记录个数的比例约为 99%,同时各种入侵活动数据记录也相应被均匀分配到这 5 个数据子集中。在实现聚类算法过程中忽略类标识属性,仅供算法结果分析。实验结果如表 1 所示。

表 1 改进的 K-means 算法的检测结果

样本集	正常记录 数量	异常记录 数量	检测率/%	误检率/%
数据子集 1	10 000	100	89.4	6.17
数据子集 2	10 000	100	92.3	5.81
数据子集 3	10 000	100	94.7	3.93
数据子集 4	10 000	100	87.9	3.46
数据子集 5	10 000	100	92.3	2.05

图 1 显示了本文算法对于不同攻击类别和正常类别数据的检测率。

通过上面的数据分析可以看到,利用该方法进行异常入侵检测可以有效地将正常数据和攻击数据区分开来,有很高的可行性和有效性。

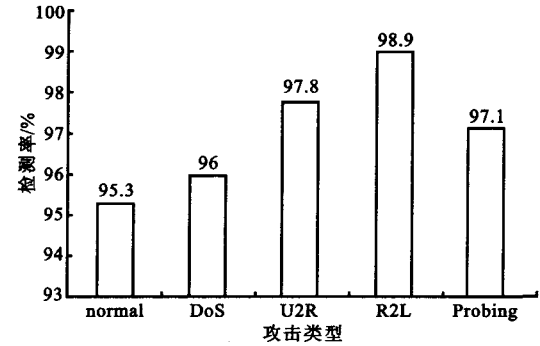


图 1 对于不同攻击类别和正常类别数据的检测率

为了对比本文算法与 K-Means 算法和 IKCM 算法^[5]对单独攻击种类的检测性能,分别选取 KDD CUP 1999 数据集四大攻击类型中的一种攻击作为代表(其中 Neptune 属于 DoS 攻击,Buffer_overflow 属于 U2R 攻击,Guess_password 属于 R2L 攻击,PortswEEP 属于 Probing 攻击),得到实验结果如表 2 所示.

表 2 本文算法与其他算法对单种攻击的检测率对比 (%)

类别	Neptune 检测率	Buffer_ overflow 检测率	Guess_ password 检测率	PortswEEP 检测率
K-Means	95	83.7	79.1	89.3
IKCM	95.3	97.2	99	95.8
本文算法	96.1	98	99.2	97.5

图 2 表示的是本文算法与传统 K-Means 算法和 IKCM 算法对四种不同攻击类型的误检率对比的折线图.

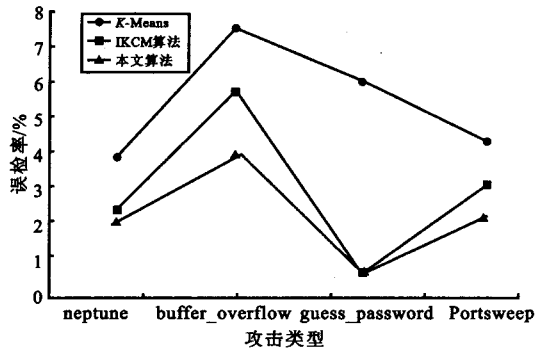


图 2 本文算法与其他算法对四种攻击类型的误检率对比

从表 2 和图 2 中可以看出,对于这四种常见攻

击类型本文算法的检测率与误检率都明显优于传统的 K-Means 算法.与 IKCM 算法相比,除 guess_password 攻击类型外,对其他三种攻击在检测率和误检率上都有提高,尤其是在误检率上,本文算法明显优于 IKCM 算法.

4 结束语

文中分析了聚类算法在异常入侵检测中的应用,对现有的 K-Means 算法进行了改进,并利用 KDD Cup1999 实验数据进行仿真实验,其结果表明该算法能有效提高检测的效率,较之 IKCM 算法在误检率上有很大提高.日后的研究方向是将该算法应用在大规模网络数据中实时地对未知入侵行为的检测.

参考文献:

[1] Lee W K, Stolfo S J, Mok K W. A data mining framework for building intrusion detection models[C]// Proceeding of the IEEE Symposium on Security and Privacy. California, IEEE Computer Society Press, 1999:120-132.

[2] Erbacher R F, Walker K L, Frincke D A. Intrusion and misuse detection in large-scale systems[J]. IEEE Computer Graphics and Applications, 2002, 22(1):38-47.

[3] Richard J Hathaway, James C Bezdek. Extending fuzzy and probability clustering to very large data sets[J]. Computational Statistics & Data Analysis, 2006, 51(1):215-234.

[4] Hansen P, Mladenovic N. J-means: a new local search heuristic for minimum sum-of-squares clustering[J]. Pattern Recognition, 2002, 34(2):405-413.

[5] Li Han, Zhang Nan, Bao Lihui. Using an improved clustering method to detect anomaly activities[J]. Wuhan University Journal of Natural Sciences, 2006, 11(6):1814-1818.

[6] University of California. KDD CUP 1999 dataset[EB/OL]. (1999-10-28)[2010-1-20]. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.

作者简介:

李涵女, (1978-), 硕士, 讲师. 研究方向为网络安全、信息处理.