

一种 K-MEANS 算法在网络异常检测中的应用

刘 涛¹, 马晓宇², 胡 景¹

(1 西安科技大学 通信学院, 陕西 西安 710054; 2 陕西高速交通工贸有限公司, 陕西 西安 710054)

摘 要: 在研究 K-MEANS 算法和网络入侵的基础上将一种已知聚类中心的 K-MEANS 聚类算法用于网络的异常检测中. 该算法避免了由于传统聚类算法随机选取初始聚类中心而带来的网络异常检测中检测率低的问题. 在实例中验证了该算法的可行性和优越性. 结果表明该算法相对传统聚类算法在检测率方面有了很大提高, 并且能通过无监督学习的方法来获得对新型攻击的检测.

关键词: 入侵检测; 异常检测; K-MEANS

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2012)05-0042-04

A K-MEANS Algorithm Applied in Network Anomaly Detection

LIU Tao¹, MA Xiao-yu², HU Jing¹

(1 School of Communication Engineering, Xi'an University of Science and Technology, Xi'an 710054, China;

2 Shaanxi Rapid Transit Industry and Trade Co., Ltd., Xi'an 710054, China)

Abstract: This paper studied a known cluster centers of k-means clustering algorithm for anomaly detection of network. The algorithm avoids the low detection rate of network anomaly detection problems for traditional clustering algorithms for random selecting initial cluster centers. Experiment demonstrated that the algorithm is feasible. Results showed that this algorithm comparing to traditional clustering algorithms in terms of detection rates have been greatly improved. Through the unsupervised learning method to obtain the detection of new attacks.

Key words: intrusion detection; anomaly detection; K-MEANS

1 引言

网络入侵检测是利用检测技术检测入侵行为. 它通过收集和分析网络行为、安全日志、审计数据、其他网络上可以获得的信息以及计算机系统中若干关键点的信息, 检查网络或系统中是否存在违反安全策略的行为和被攻击的迹象. 入侵检测可分为两类^[1]: 异常检测模型和误用检测模型. 异常检测模型用于检测与可接受行为之间的偏差; 误用检测模型用于检测与已知的不可接受行为之间的匹配程度. 目前, 异常检测技术研究已经转移到了无监督的入侵检测. 所谓无监督^[2], 即使用一组没有标记的数据作为输入, 发现其中存在的攻击数据, 无监督的入侵检测技术有一个基本的假设, 就是正常数据和异常

数据有定性的不同, 这样才可以将它们区分开来, 例如通过一般的分析, 可以知道 smurf 攻击数据在属性取值和模式上与正常数据有很大不同, 所以可以利用无监督的入侵检测技术来有效地检测 smurf 攻击. 近年来研究人员将聚类方法用于网络入侵检测系统^[3-4], 对异常行为进行检测, 提高了检测率.

本文在前人研究的基础上, 将一种基于已知聚类中心的 K-MEANS 算法应用于网络异常检测. 实验结果表明, 该算法能够在提高检测率的基础上, 较准确地发现新型攻击类型.

2 K-MEANS 算法

2.1 K-MEANS 算法

K-MEANS 算法接受输入量 k , 然后将 n 个数

收稿日期: 2011-09-11; 修回日期: 2011-10-26

基金项目: 陕西省教育厅科学研究计划(2010JK659)

据对象划分为 k 个聚类以便使得所获得的聚类满足:同一聚类中的对象相似度较高,而不同聚类中的对象相似度较小.聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”(引力中心)来进行计算的^[5].

K-MEANS 算法的工作过程是首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心;而对于所剩下其它对象,则根据它们与这些聚类中心的相似度(距离),分别将它们分配给与其最相似的(聚类中心所代表的)聚类;然后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值);不断重复这一过程直到标准测度函数开始收敛为止,一般都采用均方差作为标准测度函数. k 个聚类具有各聚类本身尽可能紧凑,而各聚类之间尽可能分开的特点.

K-MEANS 算法根据输入参数 K ,将数据集划分为 K 个簇.算法采用迭代更新的方法,在每一轮中以及 K 个参照点将其周围的点分别组成 K 个簇,而每个簇的中心将被作为下一轮迭代的参照对象.

K-MEANS 算法的处理流程为:

(1) 从 n 个数据对象任意选择 k 个对象作为初始聚类中心;

(2) 循环(3)到(4)直到每个聚类不再发生变化为止;

(3) 根据每个聚类对象的均值(中心对象),计算每个对象与这些中心对象的距离,并根据最小距离重新对相应对象进行划分;

(4) 重新计算每个(有变化)聚类的均值(中心对象).

2.2 传统 K-MEANS 算法的缺陷

传统 K-MEANS 算法的缺陷表现在中心点个数 K 的选择^[6-7]和中心点的选择上.传统 K-MEANS 算法中, K 值的选择往往是凭借经验,根据数据分布、问题性质得出的.这样做,使得聚类的结果存在一定的偏差,导致偏差的原因主要是由于 K 值选择不合理造成的. K 的取值过大,会导致聚类的结果偏多,把一些本来可以划分在同一类的点分开;而 K 的取值过小,则会导致聚类的结果偏少,把一些本来可以分开的点划分在同一类.

在初始中心点的选择上,传统的 K-MEANS 算法采取的是随机选取中心点的方法,而采用这种方法选取的初始中心点,会由于选取的中心点不同,导致每次聚类的结果不一样,使得聚类结果不稳定.因此,本文针对传统的 K-MEANS 算法的缺点和网络异常检测的特点,对传统算法做以改进.

3 基于已知聚类中心的 K-MEANS 算法的网络异常检测

3.1 基于已知聚类中心的 K-MEANS 算法

3.1.1 初始聚类中心的选取

为了使聚类结果更加准确,本文在进行聚类前,先对中心点进行初始化,使得初始的聚类中心之间的距离为数据集中相聚最远的点,从而使聚类的结果更加准确.

初始化 K 个中心点,具体方法如下:

(1) 先在数据集中选择两个距离(欧几里得距离,以下所说的距离都为欧几里得距离)最远的点,作为初始的两个中心点;

(2) 算出已经产生的中心点加权平均,得到一个临时点;

(3) 在数据集剩下的数据中,选择一点,与(2)中产生的临时点距离最远的点,作为另一个初始的中心点;

(4) 重复(2)和(3),直到选出 K 个中心点为止.

3.1.2 算法实现的改进

传统 K-MEANS 算法在做聚类操作时,通常采用递归思想.然而,使用递归算法时,往往使栈内存的资源得不到释放,很容易造成栈溢出,这样使得算法只能处理少量的数据,对于海量数据的处理则是无能为力.文中采用 while 循环结合栈来代替递归算法,从而可以避免栈溢出的问题.算法流程如图 1 所示.

3.2 基于已知聚类中心的 K-MEANS 算法用于网络异常检测

为了验证已知聚类中心的 K-MEANS 算法在网络中应用的效果,本文使用经典数据集 KDD CUP 99 进行实验,样本属性共包括了 41 个特征,其中 7 个离散属性,34 个连续属性.

在异常检测的实验中数据集必须满足 2 个前提^[6-8]:① 数据集中绝大部分的数据都是正常数据;② 入侵数据在某些属性的取值上偏离正常取值范围,这样才能使数据集模拟真正的网络数据.本次实验使用的数据样本是从 KDD CUP99 10%的数据集中提取的,并且严格遵照 2 个前提条件.对异常样本分布进行了仿真,使得训练数据集中共包含 15 类异常样本,每一类异常样本都符合 KDD CUP99 中样本的分布,然后使用训练集对该算法进行训练,获得初始聚类中心.测试样本集是从 KDD CUP99 corrected 数据集中提取的剩余 7 类异常样本,每一类

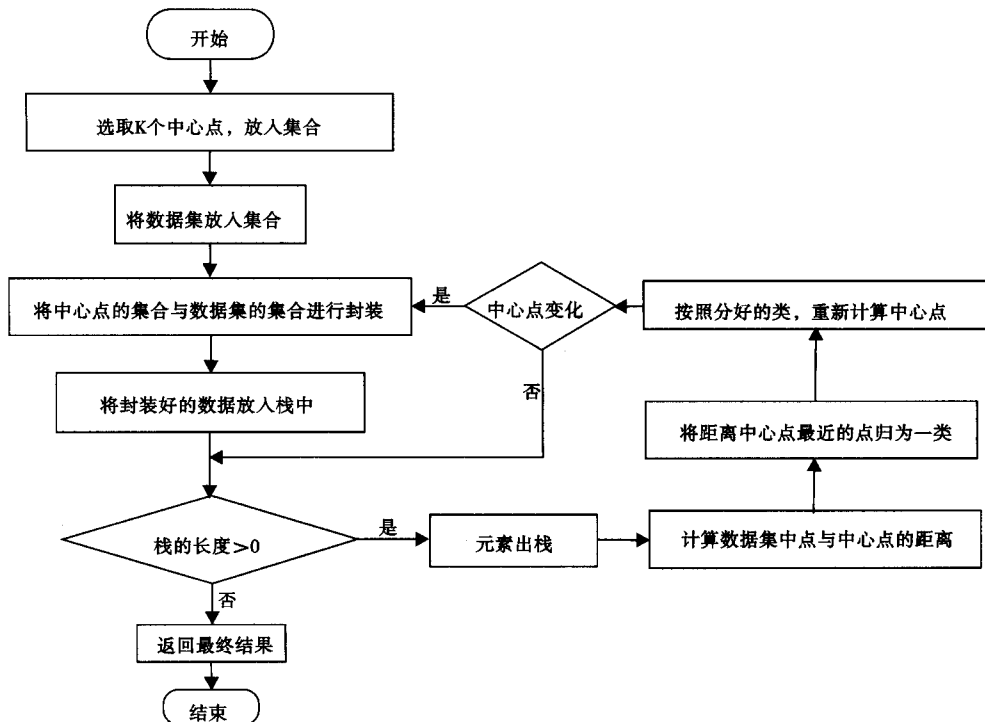


图1 基于栈实现的 K-MEANS 算法流程图

异常样本同样都符合 DKK CUP99 中样本的分布, 用不同的聚类方法和本文的聚类算法分别对这个测试样本集进行测试实验。

由于在实际应用中对象数据选用的度量单位会影响聚类分析结果, 不同度量单位可能产生迥异的聚类结构, 因此为避免对度量单位选择的依赖, 本文已先对数据样本进行了标准化和归一化处理。实验首先是在数据集样本处理完成的基础上, 将选出的训练集放入 K-MEANS 算法进行训练。训练完后, 将剩下的 7 种未知的攻击类型放入训练集进行比较, 看看能否检测出其为新的入侵。

4 实验测试及分析

4.1 实验过程

实验采用可以通过阈值的选定发现新入侵的思想, 方法如下:

(1) 将 15 种已知的攻击类型作为训练集, 使用 K-MEANS 算法找到 15 个聚类, 找出每个聚类中距离其中心点最远的点, 算出其距离中心点的距离, 作为此类攻击类型的阈值。

(2) 算出新的入侵与训练集中的每种入侵的中心点的距离, 如果该距离小于所有已知攻击类型的阈值, 则将该入侵归入距离最近的中心点所在的入侵类型。

(3) 如果该距离大于所有已知入侵类型的阈值, 则将该入侵视为新的入侵。

将传统的 K-MEANS 算法结合上述入侵检测思想进行入侵检测, 两次实验结果如表 1、表 2 所示。

表1 使用传统 K-MEANS 算法的第一次检测结果

攻击名称	总数	检测数	漏检数	检测率
warezmaster	20	14	6	0.7
perl	3	0	3	0
phf	4	2	2	0.5
pod	264	200	64	0.7575758
rootkit	10	3	7	0.3
spy	2	0	2	0
multihop	7	0	7	0
合计	310	219	91	0.7064516

表2 使用传统 K-MEANS 算法的第二次的检测结果

攻击名称	总数	检测数	漏检数	检测率
warezmaster	20	12	8	0.6
perl	3	1	2	0.3333333
phf	4	1	3	0.25
pod	264	192	72	0.7575758
rootkit	10	3	7	0.3
spy	2	1	1	0.5
multihop	7	0	7	0
合计	310	210	100	0.6774194

可以看出,在传统 K-MEANS 算法的基础上加入阈值的判定是可以检测出新型入侵的。

最后,本文的 K-MEANS 算法进行入侵检测,两次实验结果相同,如表 3、表 4 所示。

表 3 使用本文 K-MEANS 算法的第一次检测结果

攻击名称	总数	检测数	漏检数	检测率
warezmaster	20	18	2	0.9
perl	3	0	3	0
phf	4	4	0	1
pod	264	264	0	1
rootkit	10	5	5	0.5
spy	2	1	1	0.5
multihop	7	7	0	1
合计	310	299	11	0.964516129

表 4 使用本文 K-MEANS 算法的第二次检测结果

攻击名称	总数	检测数	漏检数	检测率
warezmaster	20	18	2	0.9
perl	3	0	3	0
phf	4	4	0	1
pod	264	264	0	1
rootkit	10	5	5	0.5
spy	2	1	1	0.5
multihop	7	7	0	1
合计	310	299	11	0.964516129

4.2 实验分析

本文的 K-MEANS 算法与传统的 K-MEANS 算法相比,前者 K-MEANS 算法的检测结果始终不变,而传统的 K-MEANS 算法每次的检测结果都有变化。

导致这种结果的原因主要是由于传统 K-MEANS 算法和本文 K-MEANS 算法在选取初始中心点上的不同。传统 K-MEANS 算法在选取中心点上采取的是随机的方法,导致每次选取的中心点都不一样,最终导致聚类结果不一样,从而使检测的结果不稳定。而且,由于传统 K-MEANS 算法的中心点是随机选取的,所以在聚类的时候可能会出现不同的入侵类型分不开的情况。使用本文的 K-MEANS 算法,中心点是在聚类前已初始化好的,只要训练集不变,初始中心点就不会变化。且初始化的中心点之间的距离都是最大的,因此,在聚类时,就会将不同类型的记录分得比较开,从而使聚类的结果更准确。

5 结束语

本文利用基于已知聚类中心的 K-MEANS 算法进行网络异常检测。该算法通过对已知样本训练获得初始聚类中心,不会因为聚类中心的随机选取而使聚类的效果很差;运用 while 循环结合栈的方式来代替递归算法,从而可以避免栈溢出的问题;同时算法不预设最终需要的聚类数目,而是通过聚类过程自动确定聚类数目,建立聚类类别,有利于入侵检测系统发现新型攻击。实验结果显示基于已知聚类中心的 K-MEANS 算法比传统 K-MEANS 算法在检测率方面有了提高,同时检测率稳定,并能检测到新类型攻击。

参考文献:

- [1] 程玉青,梅登华,陈龙飞. 基于数据挖掘的入侵检测系统模型[J]. 计算机技术与发展,2009,19(12): 123-124.
- [2] 向继,高能,荆继武. 聚类算法在网络入侵检测中的应用[J]. 计算机工程,2003,29(16):48-50.
- [3] 王彦涛,张凤斌. 一种改进的 K-MEANS 算法在入侵检测中的应用[J]. 电脑知识与技术,2009,34(5): 9824-9827.
- [4] Zong S, Khoshgoftaar T, Seliya N. Clustering-based network intrusion detection[J]. International Journal of Reliability, Auality and Safety Engineering, 2005(6):12-29.
- [5] 杨斌,刘卫国. 一种基于聚类的无监督异常检测方法[J]. 计算机工程与应用,2008,44(1):1-2.
- [6] Eleazar Eskin, Andrew Arnold, Michael Prerau, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data [C]// Proc of Applications of Data Mining in Computer Security. Boston: Kluwer Academic Publishers, 2002:78-99.
- [7] 叶芳,吴中福,刘勇国. 网络入侵的聚类算法研究与实现[J]. 重庆大学学报,2004,27(3): 46-52.
- [8] Srinivas Mukkamala, Andrew H Sung. Feature selection for intrusion detection with neural networks and support vector machines [J]. Journal of the Transportation Research Record, 2003, 1822(3): 33-39.

作者简介:

刘涛男,(1972—),博士.研究方向为数据挖掘、网络安全、人工智能。