改进的 K-means 算法在入侵检测中的应用

黎银环[1],张 剑[2]
(1. 江门职业技术学院,广东 江门 529000;
2. 深圳市安证计算机司法鉴定所,广东 深圳 518028)

摘 要:传统 K-means 聚类算法存在初始聚类中心选取敏感且需要预先设定聚类数等不足,导致入侵检测效率较低。为了提高入侵检测的准确性,提出一种改进的 K-means 算法。采用分离预处理记录属性的方法,在随机抽取的数据子集中基于密度距离生成初始聚类中心;利用类内最大相似度距离和类间最小相似度距离动态生成新类而无须事先确定 K 值。通过 KDDCUP99 数据集仿真实验表明,与传统的 K-means 聚类算法相比,改进的 K-means 算法有效提高了入侵检测的检测率,降低了误检率,缩短了检测时间。

关键词:入侵检测;聚类算法;K-means 算法
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2013)01-0165-04
doi:10.3969/j.issn.1673-629X.2013.01.041

Application of Improved K-means Clustering Algorithm in Intrusion Detection

LI Yin-huan[1], ZHANG Jian[2]
(1. Jiangmen Vocational and Technical College, Jiangmen 529000, China;
2. Shenzhen Detected Network Security Computer Technology Co., Ltd., Shenzhen 518028, China)

Abstract: In the traditional K-means algorithm, the initial cluster center is selected sensitively and the number of clusters must be given in advice, which leads to low efficiency in intrusion detection. In order to improve detection accuracy, an improved K-means algorithm is proposed. The method of separation pretreatment record attributes is used. In randomly selected sub-data set, the initial cluster center is generated based on the density and distance. Use the largest similarity distance in classes and between classes to dynamically generate new classes without having to predetermine value of K. Simulation experiment is done in KDDCUP99. Compared with the traditional K-means clustering algorithm, the improved algorithm improves the detection rate, reduces the false detection rate and shortens the detection time.

Key words: intrusion detection; clustering algorithm; K-means algorithm

0 引 言

入侵检测(ID)是通过收集和分析计算机网络或计算机系统中若干关键点信息,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。入侵检测软件与硬件的组合便是入侵检测系统(IDS)[1]。随着网络技术和网络规模的迅速发展,系统要收集的数据量日益增加。如何从海量数据中提取与系统安全相关的数据生成入侵检测规则和建立检测模型已成为网络安全检测的研究热点,许多学者提出了很多新的思想和算法[2-9]。Portnoy 首先提出了基于聚类分析的入侵检测技术[2]。其基本思想是入侵与正常模式的不同及正常行为数目应远大于入侵行为数目的条件,因此能够将数据集划分为不同的类别,由此分辨出正常和异常行为来检测入侵。此后,学者们对传统 K-means 聚类算法在入侵检测的应用进行了大量分析和改进。傅涛等[7]针对传统 K-means 算法的聚类结果随初始聚类中心的不同导致聚类结果不稳定,提出了基于 POS 的 K-means 算法,使聚类结果不会陷入局部最优解。谢慧[9]等针对现有的入侵检测对未知攻击检测率和误检率方面的不足,提出了基于蚁群聚类的入侵检测系统。但网络入侵随机性强,事先难以确定攻击的种类和数目,网络数据包中的属性值有连续型和离散型,传统 K-means 聚类算法在入侵检测应用中仍存在不足。

为此,文中提出一种改进的 K-means 聚类入侵检测算法。采用分离定义记录属性的方法,在抽取的数

收稿日期:2012-05-20;**修回日期**:2012-08-22
基金项目:国家发展和改革委员会信息安全专项-电子数据勘查取证服务项目(发改办高技[2009]1885)
作者简介:黎银环(1975-),女,广东江门人,讲师,硕士,CCF 会员,主要研究领域为网络与信息安全。

据子集中基于密度距离产生较好的初始聚类中心;利用类内最大相似度距离和类间最小相似度距离动态生成新类而无须事先给定准确的 K 值。仿真实验表明,与传统算法相比较,改进算法应用在入侵检测系统中能有效提高检测率和降低误检率。

1 传统 K-means 算法

1.1 传统 K-means 算法的特点

K-means 算法由 MacQueen[10] 首次提出,根据用户输入的最终分类个数 K,随机选择 K 个初始的聚类中心,通过不断地迭代计算,直到得到最终的聚类结果。

通常采用均方误差作为标准测量函数,定义如公式(1)所示:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} |x_j - m_i|^2 \tag{1}$$

其中 x_j 是空间中的一个数据点,m_i 是聚类 C_i 的均值。

传统 K-means 算法是基于爬山式的优化搜索算法,算法简洁,可伸缩性较好且快速有效。其计算复杂度为 $O(nKt)$,其中 n 为记录个数,K 为聚类个数,t 为迭代次数。资源消耗小,空间复杂度小。当聚类数据对象是密集型的,而且类与类之间区别明显时,算法的聚类效果较好。

1.2 K-means 算法在入侵检测应用中存在的不足

由于算法具有简洁、计算复杂度小等优点,将 K-means 聚类算法应用于组建无监督的入侵检测系统具有重大的研究价值。但在实际应用中还存在一些不足之处:

(1)生成的聚类数要预先给定,初始聚类中心的选取非常敏感。不准确的 K 值会导致聚类质量下降,随机选取不同的初始聚类中心点,容易导致聚类陷入局部最优。入侵检测的网络数据包通常是实时抓取的,难以预先确定聚类数,致使聚类结果稳定性和可靠性较差。

(2)传统的 K-means 算法只能处理连续型数值而不能处理离散型数值。但网络数据包中的属性值有连续型和离散型,如协议和服务名称等为字符型离散属性,须将其预处理转化为数值型。

(3)对噪声和异常点很敏感,少量的孤立点数据会对计算平均值产生很大的影响。

2 改进的 K-means * 算法

针对传统 K-means 算法在网络入侵检测应用中存在的不足,文中主要对算法的数据类型预处理、初始中心选取和 K 值确定问题进行改进。

2.1 算法相关定义

参文[8,11,12]中的相关定义及文中 K-means * 算法中的定义如下:

定义 1 告警数据库为 D,包含 n 条告警记录集 $T = \{T_1, T_2, \cdots, T_n\}$ $(n \geqslant 1)$。属性集 X 由 m 个特征属性组成,$X = \{X_1, X_2, \cdots, X_m\}$,包含数值属性子集 X_d 和字符型属性子集 X_c,满足关系:$X = X_c \cup X_d$ 且 $X_c \cap X_d = \varnothing$(空集)。每条告警记录 T_i 由 m 维属性组成 $T_i = (x_{i1}, x_{i2}, \cdots, x_{im})$。

定义 2 设 T_i 和 T_j 为任意两条告警记录,则 T_i 和 T_j 之间的数值型属性的相似度距离 $\mathrm{Sim}_d(T_i, T_j)$ 采用欧几里德距离来计算:

$$\mathrm{Sim_d}(T_i, T_j) = \left(\sum_{h=1}^{p} |x_{ih} - x_{jh}|^2\right)^{\frac{1}{2}} \tag{2}$$

(p 为数值型特征的个数,$1 \leqslant p \leqslant m$, $i \neq j$)

定义 3 设 T_i 和 T_j 为任意两条告警记录,则 T_i 和 T_j 之间的字符型属性的相似度距离 $\mathrm{Sim_c}(T_i, T_j)$ 为:

$$\mathrm{Sim_c}(T_i, T_j) = \sum_{h=1}^{q} S(x_{ih}, x_{jh}) \tag{3}$$

(q 为字符型属性的个数, $1 \leqslant q \leqslant m$, $i \neq j$)

其中 $S(x_{ih}, x_{jh})$ 表示 x_{ih} 和 x_{jh} 第 h 个字符属性的相似度距离[12],并有

$$S(x_{ih}, x_{jh}) = \begin{cases} 1, x_{ih} \neq x_{jh} \\ 0, x_{ih} = x_{jh} \end{cases} \tag{4}$$

定义 4 任意两条告警记录 T_i 和 T_j 之间的相似度距离可表示为:

$$\mathrm{Sim}(T_i, T_j) = \mathrm{Sim_c}(T_i, T_j) + \mathrm{Sim_d}(T_i, T_j) \quad (i \neq j) \tag{5}$$

定义 5 聚类集 $C = \{C_i\}$ $(i = 1, 2, \cdots, K)$。其中 $C_i = \{T_f, T_r, \cdots, T_g\}$ $(1 \leqslant f, g \leqslant n$ 且 $f \neq g$, r 为记录个数) 为包含 r 个记录的第 i 个聚类。

定义 6 C_i 类的聚类中心 m_i 表示为:

$$m_i = m_i^d + m_i^c \tag{6}$$

其中数值属性的聚类中心 m_i^d 取值为该聚类中告警记录对应该属性的算术平均值:

$$m_i^d = \frac{1}{r} \sum_{j=f}^{g} \mathrm{Sim_d}(T_i, T_j)$$

字符属性聚类中心值为聚类中告警记录对应该属性出现频率最高的属性值:$m_i^c = \mathrm{Max}(\mathrm{Sim_c}(T_i, T_j))$。

定义 7 告警记录 T_i 与当前聚类 C_j 的相似度可以用其与聚类中心 m_j 的相似度距离表示:

$$\mathrm{Sim}(T_i, m_j) = \mathrm{Sim_d}(T_i, m_j) + \mathrm{Sim_c}(T_i, m_j) \tag{7}$$

告警记录 T_i 与聚类集 C 各聚类的聚类中心间的最小相似度距离为:

$$\mathrm{Min}(\mathrm{Sim}(T_i, C)) = \mathrm{Min}(\mathrm{Sim}(T_i, C_j)) \tag{8}$$

定义 8 任意两两聚类 C_i 和 C_j 间最小相似度距离

为 SBC,可用聚类中心 m_i 和 m_j 的相似度距离其中的最小值表示。

$$SBC = Min(Sim(m_i, m_j)) \quad (i \neq j) \tag{9}$$

定义 9　包含 r 个数据对象的第 C_i 类内数据对象相似度距离的平均值 SWC_i 可表示为:

$$SWC_i = Avg(\sum_{h=1}^{r-1}\sum_{j=h+1}^{r} Sim(T_h, T_j)) \quad (h \neq j) \tag{10}$$

定义 10 任意告警记录归属到某类的最大相似度距离为:

$$SWC = Max(SWC_i) \quad (i = 1, 2, \cdots, K) \tag{11}$$

定义 11 记录分布密度函数定义如下:

$$d_i = \frac{z_i}{\sum_{j=1}^{l} z_j} \tag{12}$$

其中 $z_i = \sum_{j=1, j\neq i}^{l} \frac{1}{Sim(T_i, T_j)}$,$d_i$ 越大,样本点的密度越大,对分类的影响就越大。

2.2　初始聚类中心的确定

初始聚类中心的选择对聚类质量影响较大,选择时应该符合类中心的样本点密度较高且类中心间的相似度距离应尽可能大的要求,这可以减少算法因为初始化不理想而导致聚类结果对初值的依赖。为提高搜索初始中心的效率,考虑密度和相似度距离对初始聚类中心的影响,采用抽样方法从告警数据库 D(包含 n 条记录)中随机抽取 l 个数据子集 $D_1, D_2, \cdots, D_l$,每个子集包含 n'(l, $n' << n$)条记录,利用 Findm(D, l, n') 函数,产生 m_1, m_2, m_3 三个初始聚类中心,聚类中心靠向记录对象密集区。Findm(D, l, n')函数的实现如下:

随机抽取数据子集 $D_1, D_2, \cdots, D_l$;

For $j = 1$ to l do

{For $i = 1$ to n'

{根据公式(12)计算数据子集 D_j 各记录的分布密度 d_i ;

$m_j = Max(d_i)$; };

根据公式(6)计算 $\{m_j\}$ 的聚类中心设为 m_1}

For $j = 1$ to l do

根据公式(5)计算 Sim (m_1, m_j);

$m_2 = Max(Sim(m_1, m_j))$;

For $j = 1$ to l do

根据公式(5)计算 Sim (m_2, m_j);

$m_3 = Max((Sim(m_1, m_j) + Sim(m_2, m_j))$;

输出初始聚类中心 m_1, m_2, m_3 。

2.3　改进的 K-means * 算法

在本算法中,新聚类的生成和 K 值的确定是聚类算法的关键,其实质就是通过多次迭代计算实现类间相似度距离越小,类内相似度距离越大。在聚类的过程中动态地调整聚类数 K 值,根据类内和类间最大相似度距离进行归类,从而获得最优的聚类结果。此部分的算法流程如图 1。

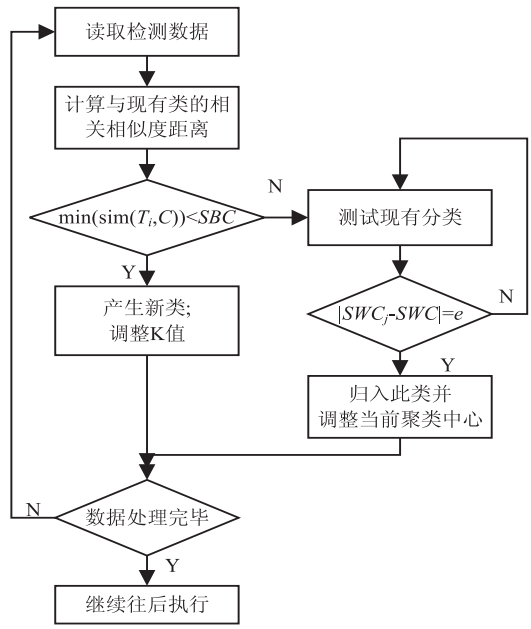

图 1　生成新类和确定 K 值流程图

改进算法的执行步骤如下:

Input:database D ;

output:{ $C_1, C_2, \cdots, C_K$ }, K

Main K-means * ( )

Begin

Findm(D, l, n');

$m = \{ m_1, m_2, m_3 \}$;

For $i = 1$ to n

{ Read (T_i);

For $j = 1$ to K

{根据公式(8)计算最小相似度距离 Min(Sim(T_i, C));

根据公式(9)计算两两聚类间的最小相似度距离值 SBC;

根据公式(10)和(11)分别计算 SWC_i 和 SWC; }

IF Min(Sim(T_i, C))<SBC　then

{ $C_{K+1} = \{T_i\}$;

$m_{K+1} = T_i$;

$m = \{m_1, m_2, \cdots, m_{K+1}\}$

$C = C_K \cup C_{K+1}$;

$K = K + 1$;}

Else

For $j = 1$ to K

If | SWC_j -SWC| $\leqslant \varepsilon$ then

{ $C_j = C_j \cup \{T_i\}$;

重新计算 m_j ;}}

Output(C);
Output (K);
End

3 仿真实验

本实验采用 KDDCUP99 入侵检测数据集作为验证数据源进行仿真实验。实验硬件平台采用 CPU 为 2.0G,内存为2GB 的个人计算机,软件平台为 Windows XP SP3,数据库为 MS SQL Server 2005,在 VC++6.0 环境下实现程序设计。10% KDDCUP99 数据集共有 494 015 条记录,包括22 种攻击,主要为4 大类网络攻击类型:DoS,Probe,U2R 和 R2L。每一条记录由 34 个数值属性和 7 个字符属性构成。其中正常数据记录占 19.68%,异常数据记录占 80.32%。为确定初始聚类中心,从数据库中随机抽取 10 个数据子集,每子集包含 1000 条记录进行多次实验。性能评估实验采用随机抽取的 3 组样本数据集作为测试数据, 每个数据子集包含 10000 条记录,异常数据记录个数占约 1.7% ~1.9%,包含不同攻击类型。数据样本如表 1。

表 1 数据样本集

样本集	正常记录数	异常记录数	攻击种类数
A	9810	190	20
B	9829	171	16
C	9819	181	18

实验中使用攻击检测率、误检率和检测时间来评价实验结果。与传统 K-means 算法对比,从表 2 可以看出,本算法的检测率平均提高约 7.68%,误检率平均降低了 2.86%;三组数据的检测时间减少约 15%,表明改进算法能有效聚类并降低了时间复杂度。

表 2 两种算法的检测结果对比

样本集	K-means			K-means *		
	检测率(%)	误检率(%)	检测时间(S)	检测率(%)	误检率(%)	检测时间(S)
A	73.52	9.22	278	84.91	6.41	235
B	78.03	9.14	256	82.7	6.17	217
C	76.11	9.01	270	83.1	6.22	229
均值	75.89	9.12	268	83.57	6.27	227

从表3 中结果显示,传统的 K-means 算法在检测 U2L 和 U2R 类攻击时检测效率比较低,本算法通过对字符属性和数值属性分别预处理计算其最大相似度距离,能较准确分析不同数据属性的特点,从而减少了聚类中的误判,提高了这两类攻击的检测效率。

表 3 两种算法对各种攻击的检测效率

攻击类型	检测率(%)			误检率(%)		
	K-means	K-means *	提高	K-means	K-means *	降低
DoS	87.18	91.91	4.73	2.01	1.86	0.15
Probe	93.12	94.32	1.20	2.81	2.21	0.60
U2R	70.01	81.62	11.61	15.18	10.45	4.73
R2L	53.02	66.24	13.22	16.76	10.39	6.37
均值	75.83	83.52	7.69	9.19	6.23	2.96

4 结束语

随着网络技术的不断发展,入侵技术和攻击手段的隐蔽性和复杂性日益明显,部分入侵行为通过伪装等手段造成算法检测困难,系统和网络中的异常记录往往蕴含着比普通数据更为重要的信息。

文中提出的改进算法在一定程度上缓解了初始聚类中心选取敏感和避免预先设定聚类数等问题,往后还需进一步对异常点和噪声进行探测和分析,从而提高检测效率和降低误检率;聚类算法在真实网络入侵检测系统应用的适应性和高效性仍有待进一步的研究。

参考文献:

[1] 卿斯汉,蒋建春,马恒太,等. 入侵检测技术研究综述[J]. 通信学报,2004(7):19-29.

[2] Portnoy L,Eskin E,Stolfo S J. Intrusion Detection with Unlabeled Data Using Clustering[C]//Proc. of the ACM Workshop on Data Mining Applied to Security. Philadelphia,USA:ACM Press,2001.

[3] 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.

[4] 舒远仲,吴文俊,陈忠贵. 改进的蚂蚁聚类入侵检测方法[J]. 计算机工程,2011,37(6):127-129.

[5] 安宇俊,龙亚星,李 炜,等. 基于改进 k-means 算法的入侵检测研究[J]. 计算机科学,2010,37(7A):211-213.

[6] 胡艳维,秦 拯,张忠志. 基于模拟退火与 K 均值聚类的入侵检测算法[J]. 计算机科学,2010,37(6):121-124.

[7] 傅 涛,孙亚民. 基于 POS 的 K-means 算法及其在网络入侵检测中的应用[J]. 计算机科学,2011,38(5):54-55.

[8] 蒋盛益,李庆华. 一种增强的 K-means 聚类算法[J]. 计算机工程与科学,2006(11):56-59.

[9] 谢 慧,吴晓平,张志刚,等. 基于蚁群聚类的入侵检测技术研究[J]. 计算机应用研究,2010,27(8):3050-3052.

[10] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proc. of Fifth Berkeley Symposium on Mathematical Statistics and Probability,Volume I. California:University of California Press,1967:281-297.

[11] 向继东. 基于数据挖掘的自适应入侵检测建模研究[D]. 武汉:武汉大学,2004.

[12] Ralambondrainy H. A Conceptual Version of the K-means Algorithm[J]. Pattern Recognition Letters,1995,16(11):1147-1157.

改进的 K-means 算法在入侵检测中的应用

作者： 黎银环， 张剑
作者单位： 黎银环(江门职业技术学院,广东 江门 529000)， 张剑(深圳市安证计算机司法鉴定所,广东 深圳 518028)

刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)