# 基于云模型和半监督聚类的入侵检测算法

李永忠 张 杰
(江苏科技大学计算机科学与工程学院 镇江 212003)

**摘 要** 针对目前网络入侵检测率低、误报率高的问题,提出了一种将云模型和半监督聚类相结合的入侵检测算法。先对聚类算法作改进,使其能够获得稳定的聚类结果。由于属性对分类贡献程度的不同,引入了云相对贴近度的概念,给出了计算属性权重的方法。以改进的聚类方法为基础建立了云模型,对属性使用动态加权和更新云模型的方法逐渐强化分类器以指导数据的分类。KDD CUP99 实验数据的仿真结果证明了该算法的有效性。

**关键词** 云模型,聚类,入侵检测,IDS
**中图法分类号** TP393.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.032

## Intrusion Detection Algorithm Based on Cluster and Cloud Model

LI Yong-zhong ZHANG Jie
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

**Abstract** A new intrusion detection algorithm based on cluster and cloud model was proposed to solve the low rate of high false alarm rate problem in network intrusion detection. Because of the different contribution of the attributes to the classification, the attributes were given based on the concept of "clouds approach degree". The cloud model was builded based on the improved cluster in the text. Using the method of dynamic weighting and the cloud model updating for the attributes gradually strengthens the classifier to guide the data classification. KDD CUP99 data set was implemented to evaluate the proposed algorithm. Experimental results prove that the method is feasible and effective.

**Keywords** Cloud model, Cluster, Intrusion detection, IDS

## 1 引言

随着计算机网络的普及和发展,网络安全越来越受到人们的重视,入侵检测技术是网络安全防御体系的关键技术之一,其目的是通过监视系统、网络流量或系统审计数据发现网络和系统的入侵行为和企图。将机器学习方法应用于入侵检测系统,能使系统具有更强的适应性、自学习性和鲁棒性,是目前入侵检测研究的一个重要方向。

为使入侵检测系统具有更好的检测性能,本文提出了一种基于云模型和半监督聚类入侵检测算法。首先对聚类算法作改进使其能够获得稳定的聚类结果,通过对聚类结果的筛选建立正常和异常云模型,将定量的属性数值转化为定性的概念。引入云相对贴近度的概念定义高维空间样本的属性的权重[1-5],在运用云模型分类时采用了属性动态加权和不断更新云模型的方法逐渐强化分类器指导数据的分类。本文方法具有较高的鲁棒性,减少了对先验知识的需求,改善了入侵检测系统的性能。

## 2 云模型理论

云模型是建立在模糊集理论和概率论的基础上,对特点结构算法形成的定性概念与其定量数值表示之间的不确定性转换模型。云模型把定性概念的模糊性和随机性结合到一起,构成定性与定量的相互映射,作为知识表示的基础[1-5]。

### 2.1 云的定义

设 $U$ 是一个用精确数值表示的定量论域,$A$ 是 $U$ 上的定性概念,若定量值 $x\in U$,并且 $x$ 是定性概念 $A$ 的一次随机出现,$x$ 对 $A$ 的确定度 $\mu(x)\in[0,1]$ 是稳定倾向的随机数。若

$$\mu:U\to[0,1],\ \forall x\in U,x\to\mu(x)$$

则 $x$ 在论域 $U$ 上的分布称为云,每个二元组 $(x,\mu(x))$ 称为一个云滴。

### 2.2 云的数字特征

云的数字特征用期望值 $Ex$,熵 $En$ 和超熵 $He$ 来表示,它们把客观世界中的事物或人类知识中概念的随机性和模糊性集成到一起,构成定量到定性互相间的映射。

期望 $Ex$ 是在论域空间中能够代表这个定性概念的点,反映了相应模糊概念的信息中心值。

熵 $En$ 是定性概念的不确定性的度量,由定性概念的随机性和模糊性共同决定的,一般熵越大概念的随机性越宏观,表现了随机性和模糊性的关联。

超熵 $He$ 是熵的不确定性度量,是熵的熵,反映了论域空间中代表该概念所有云滴的凝聚性,即云滴凝聚的紧密度。

### 2.3 逆向云发生器

逆向云发生器是实现定量数值与定性语言之间的不确定性转换模型,它将一定数量的精确数据有效地转换为以恰当定性语言值 $Ex,En,He$ 表示的概念。本文采用文献[6]中精度较高的 $X$ 信息逆向云算法。

到稿日期:2014-08-24 返修日期:2014-09-27 本文受江苏省高校自然科学基金资助项目(05KJD52006,13KJD52004)资助。
李永忠(1961—),男,教授,硕士生导师,主要研究方向为网络安全、计算机应用、藏文信息处理,E-mail:liyongzhong61@163.com;张 杰(1988—),硕士生,主要研究方向为网络与信息安全。

$X$ 信息的一维逆向云算法，具体步骤如下：

输入：N 个云滴 $x_i$

输出：这 N 个云滴所代表的定性概念的期望 Ex，熵 En，超熵 He

(1)计算输入样本的均值 $\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_i$，一阶样本绝对中心距 $\frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{X}|$，样本方差 $S^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{X})^2$；

(2)$Ex=\bar{X}$；

(3)$En=\sqrt{\pi/2}*\frac{1}{N}|X_i-Ex|$；

(4)$He=\sqrt{S^2-En^2}$。

## 3 半监督聚类

### 3.1 聚类原理

聚类是将一个数据集划分为多个类或簇，并且使得同一簇的数据对象具有较高的相似度，而不同的簇中的数据对象不具相似性。聚类的目的是使得同一簇的数据的相似性最大，而不同簇之间的相异度最大[7-10]。在聚类的划分中，比较流行的是使用 K-means 算法，当然 K-means 也是一种常用的异常检测中的聚类分析方法，但经典的 K-means 算法本身存在着一些不足，本文对 K-means 算法作了改进[11]。

### 3.2 半监督聚类算法

半监督学习(semi-supervised learning)是一种介于无监督学习和监督学习之间的方法，其学习过程中使用的数据集通常含有少量的已经标记信息，通过这些标记信息的样本来约束指导未知样本的学习。半监督聚类是一种新的聚类方法，它综合了监督学习和无监督学习的优点，利用少量标记数据改善聚类的效果[7-11]。

在 K-means 聚类算法中主要存在的问题是初始聚类中心的选择问题和空聚类等，算法的稳定性差，本文对 K-means 作了改进，采用半监督聚类，算法步骤如下：

输入：参数 k，标记数据集 $S_l$，未标记数据集 $S_u$

输出：k 个簇

(1)利用 $S_l$ 中的标记数据确定 L 个初始聚类质心；

(2)$\forall x\in S_u$，计算其与各个聚类质心的最小距离，取最小距离的极大值对应的数据点作为下一个聚类的质心，记为第 L+1 个质心；

(3)$\forall x\in S_u$，计算其与各个聚类质心的距离，将 x 分配到与之距离最小的质心所属的簇中，更新各个聚类簇的质心；

(4)如果聚类质心为 k，重复分配 $S_l$ 和 $S_u$ 中的每个数据点到与之聚类距离最小聚类质心所属的簇中，重新计算各个聚类簇的质心，否则转向步骤(2)；

(5)对这 k 个聚类质心进行聚类，直到聚类质心不再发生变化为止；

(6)输出 k 个簇。

本算法在选取聚类中心时，采用先聚再找聚类中心的方法，增强了系统的稳定性，提高了质心聚类时的收敛速度，能够获得较高的效率。

## 4 基于云模型和半监督聚类入侵检测算法

基于云模型的入侵检测方法一般是利用云的定性推理将安全专家用自然语言所表达的定性检测规则转化为计算机能够处理的定量规则，这类方法缺乏事实依据和标准。文献[4]采用逆向云发生器从真实训练集中得到云的数字特征，形成判断规则，实现正常建模。这种方法在实际运用时，需要大量的训练数据和训练时间，并且由训练数据获得的云数字特征值并不能反映实际入侵时的情况，文章中对属性权重的计算主观性太强，在检测时阈值的设定非常困难，实用性不强。

本文算法首先对数据集使用半监督聚类算法，按聚类结果簇的大小排序选出确定的正常簇和异常簇，利用簇中的数据对象建立正常云模型和异常云模型，然后用云模型分类器对剩余的数据对象进行分类，分类结束后将其加入到对应的簇中更新云模型，重新计算各属性权重指导其它数据分类。

属性加权设定：本文参照文献[5]中的云相对贴近度的概念，设在论域空间 $U$ 中有两朵云 $A_1(Ex1,En1,He1)$，$A2(Ex2,En2,He2)$，定义 $D_{1,2}=|Ex1-Ex2|$，那么 $D_{1,2}$ 则反映了这两朵云的相对贴近度。在入侵检测过程中，如果设正常云为 $A_1$，异常云为 $A_2$，则在对每一维属性建模时，$D_{1,2d}$ 的大小反映该属性在分类中相对重要程度。用该方法对属性加权符合人们对事物概念的认知，并且动态加权能够充分利用数据的信息使得加权过程更加科学。

基于云模型和半监督聚类的入侵检测算法步骤如下：

输入：包含 n 个 d 维的数据的数据集 S，$S=S_l\cup S_u$(标记数据集 $S_l$，未标记数据集 $S_u$)

输出：数据 $x\in S_u$ 的数据类型(正常或异常)

(1)对数据集 S 使用文中 3.2 节的半监督聚类算法进行聚类处理；

(2)对聚类结果按簇的大小进行升序排列；

(3)结合数据的标记信息筛选出初始的正常簇 $C_n$ 和异常簇 $C_a$，其余数据分配到 $C_r$ 中；

(4)对 $C_n$ 中的每一维的数据利用逆向云发生器得到相应的云数字特征值 $(Ex1_i,En1_i,He1_i)$，$i=1,\cdots,d$；

(5)对 $C_a$ 中的每一维的数据利用逆向云发生器得到相应的云数字特征值 $(Ex2_i,En2_i,He2_i)$，$i=1,\cdots,d$；

(6)利用式(1)计算各个属性的权重：

$$w_i=|Ex1_i-Ex2_i|/\sum_{j=1}^{d}|Ex1_j-Ex2_j| \quad (1)$$

(7)依次从 $C_r$ 中取出一个数据对象 x，根据 X 条件正向云发生器利用式(2)计算得到异常和正常的云分类模型：

$$\mu_j=\sum_{i=1}^{d}w_i\cdot\exp[-(x-Exj_i)/2\cdot Enj_i],j=1,2 \quad (2)$$

若 $\mu_1>\mu_2$，则 x 属于正常类，将其分配给 $C_n$ 中，返回步骤(4)，更新正常云模型后转到步骤(6)重新计算各个属性的权重；否则将 $x$ 分配给 $C_a$，返回步骤(5)，更新异常云模型后再转到步骤(6)重新计算各个属性的权重，直至所有数据分类结束。

## 5 实验与分析

采用 KDD CUP 1999 数据集[12]来验证本文算法的有效性，该数据集包含了 4 种主要攻击类型：(1)拒绝服务(DoS)；(2)未经授权的远程访问(R2L)；(3)对本地超级用户的非法访问(U2R)；(4)扫描与探查(Probe)。由于 KDD 原始数据集过于庞大，针对基于云模型和半监督聚类入侵检测算法实验仿真，我们从“kddcup. newtestdata_10_percent_corrected”中另选了 3 组数据集做测试，同时也选取了其中的 500 条数据作为标识数据记录，测试数据类型及分布如表 1 所列。

表 1 实验测试数据表

| 测试数据 | 数量 | DoS(%) | R2L(%) | U2R(%) | Probe(%) |
|---|---|---|---|---|---|
| 数据集 1 | 5844 | 3.56 | 0.87 | 0.55 | 1.18 |
| 数据集 2 | 5836 | 3.72 | 0.87 | 0.46 | 1.18 |
| 数据集 3 | 5737 | 3.78 | 0.89 | 0.45 | 1.29 |

其中，DoS 的攻击为 smurf 和 neptune，R2L 的攻击为 guess_passwd，U2R 的攻击为 buffer_overflow、landmodule、perl 和 rootkit，Probe 的攻击为 portsweep。

实验采用检测率和误报率作为算法性能的度量标准。检测率＝检测出的攻击数/攻击总数；误报率＝被误报为入侵的

正常样本数/正常样本数。

半监督算法中不同的聚类质心的个数对聚类的效果影响是不同的，实验选取了不同的 $K$ 值对上述 3 组数据分别进行了测试，取它们的平均值作为检测结果。图 1、图 2 给出了不同 $K$ 值情况下云模型半监督聚类算法的检测结果。

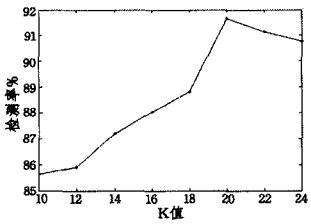


图 1　不同 K 值下的检测率

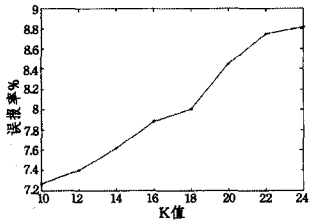


图 2　不同 K 值下的误报率

从实验结果可以看出，当 $K$ 的值逐渐增大时，误报率也随之增大，但在 $K$ 取 20 时，检测率获得最大，由此可知，$K$ 取 20 时，基于云模型半监督聚类的算法可以获得较好的入侵检测效果，其检测率达到 91.67%，误报率为 8.45%。

在基于云模型的半监督聚类算法中，$K$ 取不同值时的检测率与误报率与 K-means 算法的对比如图 3、图 4 所示。

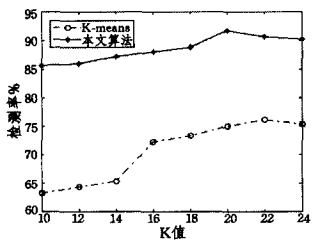


图 3　不同 K 值下检测率对比

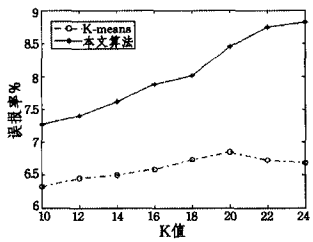


图 4　不同 K 值下误报率对比

由图 3、图 4 可以看出，在不同的 $K$ 值情况下，本文算法在检测率方面明显高于 K-means 算法，误报率与 K-means 相比显得略微偏高，但是这种误报率在可接受的范围内。

基于云模型的半监督聚类算法检测结果与一般聚类算法以及普通云模型分类器[10]的比较结果如图 5 所示。

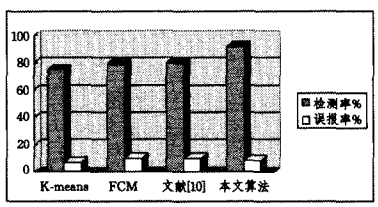


图 5　检测结果对比

图 5 给出了本文算法和其他几种算法的检测结果之间的比较，通过结果可以发现，本文算法的检测率明显高于其他 3 种算法，误报率比 K-means 略高，比其它两种算法的误报率低，证明了本文的算法具有优越的性能。

**结束语**　本文提出了一种基于云模型和半监督聚类的入侵检测算法，首先用改进的半监督的聚类对数据进行聚类处理，根据结果建立云模型，引入了云相对贴近度的概念，定义了高维空间样本在分类过程中的属性权重。在分类过程中对所建立的云模型更新和对属性实现动态加权不但能准确地反映实际数据信息而且指导了数据的分类，避免了对数据先验知识的依赖，在一定的程度上也丰富了云分类器相关的内容。实验证明了本文算法在入侵检测方面的可行性和有效性，但是本文算法的误报率仍然偏高，需要今后进一步的研究和改进。

## 参 考 文 献


[1] 李德毅，邸凯昌，李德仁，等. 用语言云模型发掘关联规则[J]. 软件学报，2000，11(2)：143-158

[2] 李德毅，史雪梅，孟海军. 隶属云和隶属云发生器[J]. 计算机研究和发展，1995，6(32)：15-20

[3] 李德毅，刘常昱. 论正态云模型的普适性[J]. 中国工程科学，2004，6(8)：28-34

[4] 吕辉军，王晔，李德毅，等. 逆向云在定性评价中的应用[J]. 计算机学报，2003，26(8)：1009-1014

[5] 付斌，李道国，王慕快. 云模型研究的回顾与展望[J]. 计算机应用研究，2011，28(2)：420-425

[6] 刘常昱，冯芒，李德毅，等. 基于云 X 信息的逆向云新算法[J]. 系统仿真学报，2004，16(11)：2417-2410

[7] Basu S, Baneriee A, Mooney R. Semi-supervised clustering by seeding[C]// Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2002: 19-26

[8] Flanagan J A. Unsupervised clustering of symbol strings[C]// International Joint Conference on Neural Networks (IJCNN'03). Portland Oregon, USA: 2003, 3250-3255

[9] Li Yong-zhong, Li Zheng-jie. Anomaly Intrusion Detection Method Based on K-means Clustering Algorithm with Particle Swarm Optimization [C]// International Conference of Information Technology, Computer Engineering and Management Sciences(ICM 2011). 2006: 415-426

[10] 姜伟，高知新，李本喜. 基于多维云模型的入侵检测[J]. 计算机工程，2006，32(24)：155-156

[11] 李涵. 基于聚类的异常检测方法的研究与实现[J]. 北京信息科技大学学报，2010，25(3)：80-83

[12] KDD CUP 1999 Data set[OL]. http://kdd.ics.uci.edu/databases/ kddcup99


(上接第 117 页)


[9] Luo X, Chan E W W, Chang R K C. Cloak: A Ten-Fold Way for Reliable Covert Communications[C]// Proc. European Symp. Research in Computer Security. Sept. 2007

[10] Luo Xia-pu, et al. TCP covert timing channels: Design and detection[C]// IEEE International Conference on Dependable Systems and Networks With FTCS and DCC. 2008: 420-429

[11] Peng P, Ning P, Reeves D. On the Secrecy of Timing-Based Active Watermarking Trace-Back Techniques[C]// Proc. IEEE Symp. Security and Privacy. May 2006

[12] Berk V, Giani A, Cybenko G. Detection of covert channel encoding in network packet delays[R]. Technical Report, TR2005 536. Department of Computer Science, Dartmouth College, 2005: 1-11

[13] Gianvecchio S, Wang H N. Detecting covert timing channels: An entropy-based approach[C]// Proc. of the 14th ACM Conf. on Computer and Communications Security. 2007: 307-316

[14] Takens F. Detecting strange attractors in turbulence: Dynamical systems and turbulence[C]// Rand D A, Young L S, eds. Lecture Notes in Mathematics. 1981, 366

[15] Eckmann J P, Kamphorst S O, Ruelle D. Recurrence Plots of Dynamical Systems [J]. Europhysics Letter, 1987, 4: 973-977

[16] Groth A. Visualization of coupling in time series by order recurrence plots[J]. Phys. Rev. E, 2005, 72: 046220

[17] Waikato VIII[OL]. http://wand.net.nz/wits/waikato/8