doi:10.3969/j.issn.1007-855x.2013.04.010

云模型半监督聚类动态加权的入侵检测方法

张 杰, 李永忠
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212003)

摘要: 针对入侵检测系统存在的检测率较低和误报率较高的问题,提出了一种云模型半监督聚类动态加权的入侵检测方法.由于属性对分类贡献程度不同,文中引入云相对贴近度的概念,给出了计算属性权重的方法.以半监督聚类算法为基础建立云模型并构造云分类器,分类时对属性使用动态加权通过对云模型的更新逐渐强化云分类器指导数据的分类.最后仿真实验表明该方法具有较好的检测效果,改善了入侵检测系统的性能.

关键词: 云模型; 半监督聚类; 入侵检测
中图分类号: TP393.4 **文献标识码:** A **文章编号:** 1007-855X(2013)04-0044-04

Dynamic Weighted Intrusion Detection Method Based on Cloud Model and Semi-Supervised Clustering

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

Abstract: In view of the low detection rate and high false alarm rate in the intrusion detection system, a new method for the intrusion detection based on cloud model and semi-supervised clustering is proposed. Due to the different contribution of the attributes to the classification, the concept of "clouds approach degree" is introduced to weight the attributes. A cloud model and classifier is established based on the semi-supervised clustering algorithm. The method of dynamic weighting is then adopted in the classification. The cloud model is updated to gradually strengthen the classifier to guide the data classification. Finally, the simulation results show that the performance of intrusion detection system is improved.

Key words: cloud model; semi-supervised clustering; intrusion detection

0 引 言

随着计算机的普及和网络的广泛应用,网络被入侵的风险性和机会也越来越多,网络安全已经成为了全球性的问题[1].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.目前,大量的入侵检测方法也相继出现.

为了提高入侵检测系统的检测性能,本文提出了一种云模型半监督聚类动态加权的入侵检测方法.本文首先通过半监督聚类的方法获得聚类结果后,再根据其中少量的标记信息筛选并建立初始的正常和异常云模型,将定量的属性数值转化为定性的概念,在定义了高维空间样本属性的权重时引入了云相对贴近度的概念,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云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.

收稿日期: 2013-05-07. **基金项目:** 江苏省高校自然科学基金项目(05KJD52006);江苏科技大学科研项目(2005DX006J).
作者简介: 张杰(1988-),男,硕士研究生.主要研究方向:网络与信息安全. **E-mail:** zxuanyi_1988@163.com
通信作者: 李永忠(1961-)男,硕士,教授,主要研究方向:网络安全,计算机应用,藏文信息处理等.
E-mail: liyongzhong61@163.com

1 云模型理论

1.1 云模型

云模型是在模糊集理论和概率论的基础上，通过特定结构算法形成的定性概念与其定量数值表示之间的转换模型. 云模型把定性概念的模糊性和随机性结合到一起，构成定性和定量相互映射，作为知识表示的基础[2].

1.2 云的定义

设 U 是一个用精确数值表示的论域，U 上对应的定性概念 A，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

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

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].

1.3 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，三个数值反映了定性概念的定量特性，构成定量到定性互相间的映射[2].

期望 Ex ：它是在论域空间中最能代表这个定性概念的值，通常是云重心对应的 x 值，Ex 反映了相应定性知识的信息中心值[3-4].

熵 En ：熵是定性概念的不确定性的度量，它是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，体现了定性概念的亦此亦彼性的裕度[3-4].

超熵 He ：超熵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凝聚的紧密度[3-4].

1.4 逆向云发生器

逆向云发生器是实现定量数值到定性语言概念的转换模型，它可以将一定数量的精确数据有效地转换为以云的数字特征 Ex,En,He 表示的定性概念. 本文采用的是文献[5]中 X 信息逆向云算法.

X 信息的一维逆向云算法，具体步骤如下：

输入：N 个样本点 x_i，其中 $i=1,2,\cdots,N$.

输出：这 N 个样本点所代表的定性概念的期望 Ex，熵 En，超熵 He.

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

2）$Ex=\overline{X}$；

3）$En=\sqrt{\pi/2}\cdot\frac{1}{N}\sum_{i=1}^{N}|x_i-\overline{X}|$；

4）$He=\sqrt{S^2-En^2}$.

1.5 *X* 条件正向云发生器

所谓云正向发生器是实现定性概念到其定量表示之间的不确定性的转换. X 条件正向云发生器指在给定的论域空间中，已知云的数字特征 Ex,En,He，如果有特定的条件 x_i，输出带有确定度 $\mu(x_i)$ 的云滴. X 条件正向云发生器算法具体步骤如下[9]：

1）生成以 En 为期望，He 为方差的正态随机数 En'，$En'=\mathrm{normrand}(En,He)$；

2）计算 $\mu(x_i)=\exp\left[-\frac{(x_i-Ex)^2}{2}\cdot En'\right]$，令 $(x_i,\mu(x_i))$，为云滴.

2 半监督聚类

2.1 聚类原理

聚类是将一个数据集划分成若干个聚类的过程，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有相似性. 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[6-7]. 聚类一直是机器学习领域的研究热点,也是数据挖掘、模式识别等领域最为常用的技术之一.

2.2 半监督聚类及算法描述

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特点,利用少量标记数据改善聚类的质量[8],提高了聚类算法收敛速度和准确性并且使系统具有较高的鲁棒性. 本文采用的半监督聚类算法流程如下:

输入:簇的数目 k,标记数据集 S_l,未标记数据集 S_u;

输出:k 个聚类簇.

步骤如下:

1) 利用 S_l 中的标记数据确定 L 个初始聚类中心;

2) $\forall x \in S_u$ 计算与各个聚类中心的最小距离,取最小距离的极大值对应的数据点作为下一个聚类的中心,记为第 $L+1$ 个中心;

3) $\forall x \in S_u$ 计算与各个聚类中心的距离,将 x 分配到与之距离最小的中心所属的簇中,重新计算各个聚类簇的中心;

4) 如果聚类中心为 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中心所属的簇中,重新计算各个聚类簇的中心,否则转向步骤2);

5) 对这 k 个聚类中心进行聚类,直到聚类中心不再发生变化为止;

6) 输出 k 个聚类簇.

3 云模型半监督聚类的入侵检测

基于云模型的入侵检测方法一般是利用云的定性推理将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再通过规则发生器实现,此类方法缺乏事实依据和标准. 文献[9]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,并且文章中对属性权重的计算主观性太强,同时在检测时阈值的取定非常困难.

本文则首先对数据集使用半监督聚类算法,对聚类结果按簇的大小排序同时根据标记信息选出初始的正常簇和异常簇,利用簇中的数据建立正常云模型和异常云模型,用所建立云模型分类器对剩余数据对象进行分类,分类同时采用更新云模型和重新计算各属性权重方法指导数据的分类.

属性权重的设定:本文参照了文献[10]中的云相对贴近度的概念,即设在论域空间 U 中有两朵云 $A1(Ex1,En1,He1)$,$A2(Ex2,En2,He2)$,定义 $D1,2 = |Ex1 - Ex2|$,那么 $D1,2$ 则反映了为这两朵云的相对贴近度. 假设在入侵检测中正常云为 $A1$,异常云为 $A2$,则在对每一维属性建立云模型时 $D1,2$ 大小则反映该属性在分类过程中相对重要程度. 用该方法对属性加权符合人们对事物概念的认知,并且采用动态加权的方式能够充分利用数据本身隐含的信息,加权方式更为科学.

基于云模型半监督聚类的入侵检测步骤如下:

输入:包含 n 个 d 维数据的数据集 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 使用本文2.2中半监督聚类算法进行聚类处理;

2) 对聚类结果按簇大小进行升序排列;

3) 结合数据的标记信息筛选出初始的正常簇和异常簇分别为 C_n 和 C_u,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_u 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) 利用公式(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公式(2)计算得到异常和正常的云分类模型

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

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_u，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.

4 实验与分析

为了验证本文方法的有效性，实验采用了 KDD CUP 1999 数据集[11]. 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). 实验选取 DoS 的攻击为 neptune 和 smurf，R2L 的攻击为 guess_passwd，U2R 的攻击为 land-module、buffer_overflow、perl 和 rootkit，Probe 的攻击为 portsweep.

为了评价入侵检测方法的性能，实验采用检测率和误报率作为算法性能的度量标准，其定义如下所示：

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

本文共做了三组实验测试算法的性能，三组实验选取了不同的数据集对算法进行了测试，其数据类型及分布如表 1 所示.

表 1　实验测试数据表

Tab. 1　Dataset of experimental test

测试数据	数量	DoS /%	R2L /%	U2R /%	Probe /%
数据集 1	5 860	3. 60	0. 89	0. 55	1. 21
数据集 2	5 852	3. 74	0. 91	0. 51	1. 23
数据集 3	5 848	3. 71	0. 87	0. 48	1. 27

实验中 k 取 25，对每个数据集进行反复实验，最终获得的平均测试结果如表 2 所示.

表 2　测试结果

Tab. 2　Test results

测试数据	检测率/%	误报率/%
数据集 1	90. 71	8. 03
数据集 2	89. 84	8. 00
数据集 3	89. 12	7. 98
平均检测结果	89. 89	8. 00

检测结果与一般聚类算法和普通云分类器的比较结果如表 3 所示.

由上述实验结果可知，本文的方法相对于普通的聚类算法和文献[12]性能上有了较大的提高，一定程度上解决了目前入侵检测存在的一些问题，但是误报率仍然偏高，对分布特殊的数据不能获得很好的分类效果等问题还需要进一步的改进.

表 3　检测结果对比

Tab. 3　Comparison of the detection results

算法	K - means/%	FCM/%	文献[12] /%	本文算法/%
检测率	74. 92	79. 25	80	89. 89
误报率	6. 85	9. 78	10	8. 00

5 结　论

提出了一种云模型半监督聚类动态加权的入侵检测方法，方法首先用半监督聚类算法对数据处理后，根据结果结合标记信息筛选数据建立初始的云模型. 将云模型难以处理高维数据的问题转化到属性权重上，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不断更新并对属性实现了动态加权，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过度依赖，在一定的程度上也丰富了云分类器的相关内容. 实验证明了该方法在入侵检测方面的可行性与有效性，在一定程度上解决了部分入侵检测算法存在的检测率低误报率高的问题，但是本文的误报率仍然偏高，有待今后进一步的研究和改进.

(下转第 59 页)