基于模糊C-均值聚类算法的入侵检测

罗军生,李永忠,杜 晓
(江苏科技大学 电子信息学院,江苏 镇江 212003)

摘 要:聚类分析是一种有效的异常入侵检测方法,可用以在网络数据集中区分正常流量和异常流量。文中采用模糊C-均值聚类算法对网络流量样本集进行划分,从中区分正常流量和异常流量,并针对入侵检测问题的特性提出了新的相似性度量方法。最后,利用KDD99数据集进行实验,证明该算法能够有效地发现异常流量。

关键词:模糊聚类;入侵检测;距离测度;混合属性;数据挖掘
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2008)01-0178-03

Intrusion Detection Based on Fuzzy C-Means Clustering Algorithm

LUO Jun-sheng, LI Yong-zhong, DU Xiao
(Sch. of Electronic Info., Jiangsu Univ. of Sci. and Tech., Zhenjiang 212003, China)

Abstract: Clustering is an effective method of anomaly intrusion detection. It could distinguish normal flow and abnormal flow in the network data set. Uses fuzzy C-means clustering algorithm for dividing the network data set into normal flow and abnormal flow. A new distance measurement method which is designed for intrusion detecting problem specially has been put forward in this paper. In the end, use KDD 1999 data set to experiment algorithm, and the result proves that this algorithm could discover the abnormal flows effectively.

Key words: fuzzy clustering; intrusion detection; distance measurement; mixed attributes; data mining

0 引 言

网络安全问题是创建信息化社会过程中所需解决的关键问题之一。随着人们对网络依赖程度的日益加深,建立行之有效的网络安全基础设施已迫在眉睫。

入侵检测系统(IDS)是网络安全基础设施的重要一环[1,2],当前的研究热点是异常检测技术。异常检测技术的最大优势在于它可以发现未知的新型攻击,这在日益复杂的网络环境中是较为理想的策略。

聚类分析是一种重要的异常检测方法。聚类分析是利用某种相似性度量,把一个未知类别的样本集组织成若干有意义的子集,要求相似度较高的样本尽量归为一类,而不相似的或相似度较小的样本则在不同的类中[3]。通过这样的划分,可将网络流量样本集中的正常流量和异常流量区分开来[4]。当前的聚类算法大多采用距离作为样本间的相似性度量。这是一种样本间的模糊关系,反映样本间的相似程度。

文中采用模糊C-均值聚类算法对KDD 99的网络流量样本集进行聚类分析。试验证明该算法可以有效应用于异常入侵检测。

收稿日期:2007-03-27
基金项目:江苏省教育资助项目(2005DX006J)
作者简介:罗军生(1980-),男,河南平顶山人,硕士研究生,研究方向为网络与信息安全、智能信息处理;李永忠,教授,研究方向为网络与信息安全。

1 基于模糊C-均值的聚类算法

1.1 数据集划分

给定 $X=\{x_1,x_2,\cdots,x_n\}$ 为模式空间中 n 个模式的一组有限观测样本集,$x_k=(x_{k1},x_{k2},\cdots,x_{kn})$ 为观测样本 x_k 的特征矢量,对应特征空间中的一个点,x_{kj} 为 x_k 的第 j 个属性的值。对给定样本集 X 的聚类分析就是要产生 X 的 c 划分。利用隶属函数 $\mu_{ik}=\mu_{xi}(x_k)$ 表示样本 x_k 与子集 X_i 的隶属关系,显然 $\mu_{ik}\in[0,1]$。且

$$M_{fc}=\{U\in R^{cn}\mid \forall i,k,\mu_{ik}\in[0,1];\forall k,\sum_{i=1}^{c}\mu_{ik}=1;\forall i,0<\sum_{k=1}^{n}\mu_{ik}<n\}$$

定义模糊聚类分析的目标函数为:

$$\begin{cases}Jm(U,P)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2\\ \text{s.t. } U\in M_{fc}\end{cases}$$

其中,U 为样本空间,P 为聚类原型,d_{ik} 为距离度量。

对于入侵检测问题,U 即为网络流量样本集[5]。

1.2 相似性度量的选取

聚类分析的前提是选取合适的相似性度量方法。

现有的聚类技术大都以"距离"作为相似性度量。由于检测对象——网络流量通常是包含混合型属性的样本,故给出如下的基于属性值范围的加权相似性度量方法[3]。

设样本集的容量为 n,其中的样本为 $x_1, x_2, \cdots, x_n$;该样本集的维数为 m,其中包含 r 个连续性属性 $a_1, a_2, \cdots, a_r$ 和 s 个离散性属性 $b_1, b_2, \cdots, b_s$;另设 x_{if} 表示第 i 个样本的第 f 个属性值。混合类型数据集的距离测度应分两部分分别计算。

1) 连续型属性。

对于连续型属性,一般以欧氏距离作为距离测度,但应该意识到两点:

(1) 单位对距离测度的结果影响很大,当一个属性的单位过小时,它对距离的影响会放大,因此,应使得属性值与单位无关,即"数据的标准化";

(2) 当属性差相同时,范围较大的属性应比范围较小的属性具有更高的权重。因此,以加权的欧氏距离作为连续型属性的距离测度。

① 首先,将数据进行标准化。

设 m_f 表示属性 a_f 的平均值,$m_f = \frac{1}{n}\sum_{k=1}^{n} x_{kf}$

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|)$$

标准化后的属性值:$z_{if} = \frac{x_{if} - m_f}{s_f}$

② 然后,计算加权的距离测度。

设 $\text{range}_1, \text{range}_2, \cdots, \text{range}_r$ 分别是 $a_1, a_2, \cdots, a_r$ 的取值范围。

$$cd(i,j) = \sqrt{\omega_1(z_{i1} - z_{j1})^2 + \omega_2(z_{i2} - z_{j2})^2 + \cdots + \omega_r(z_{ir} - z_{jr})^2}$$

其中,$\omega_f = \frac{\text{range}_f}{\sum_{k=1}^{r}\text{range}_k}$。

③ 最后,对距离进行归一化处理,即将 $cd(i,j)$ 的取值范围控制在[0,1]。

$$cd'(i,j) = \frac{cd(i,j)}{\max\{cd(i,j)\}}$$

2) 离散型属性。

对于离散型属性,首先应对各个可能的状态进行编码。而离散型属性的相似度则不能简单地以欧氏距离计算,因为编码的距离反映不出各状态间的距离。对于离散型属性只有相等或不等的关系。还应注意,和连续型属性中讨论的一样,状态较多的属性应具有更高的权重。

设 $t_1, t_2, \cdots, t_s$ 分别是 $b_1, b_2, \cdots, b_s$ 的状态数。于是,

$$dd(i,j) = 1 - \sum_{e=1}^{s}\omega_e\mu_e$$

其中,$\omega_e = \frac{t_e}{\sum_{k=1}^{s} t_k}$,$\mu_e = \begin{cases} 0, x_{ie} \neq x_{je} \\ 1, x_{ie} = x_{je} \end{cases}$

3) 合并两距离。

$$D(i,j) = \frac{r}{r+s}cd'(i,j) + \frac{s}{r+s}dd(i,j)$$

注意,$D(i,j)$ 是两数据对象 x_i, x_j 的距离,间接地反映相似度,相似度应为:

$$\text{Sim}(i,j) = 1 - D(i,j)$$

1.3 算　法

对于入侵检测聚类问题,设类别数为 c,以前 c 个样本作为每个类别的聚类原型。分别求聚类隶属度,然后根据隶属度更新聚类原型模式直到原型不再变动。并设置迭代计数器 $b = 0$。

下面给出算法的具体步骤:

step1:初始化聚类原型模式 $\boldsymbol{P}$。

step2:利用式(1) 计算隶属度值

$$u_{ik} = \left\{\sum_{j=1}^{c}\left[\left(\frac{d_{ik}^b}{d_{jk}^b}\right)^{\frac{2}{m-1}}\right]\right\}^{-1} \tag{1}$$

如果 $\exists i, r$,使得 $d_{ik}^b = 0$,则有

$\mu_{ir}^b = 1$,且 $\mu_{ir}^b = 0$,当 $i \neq r$

step3:利用式(2) 更新聚类原型模式:

$$p_i^{(b+1)} = \frac{\sum_{k=1}^{n}(u_{ik}^{b+1})^m \cdot x_k}{\sum_{k=1}^{n}(u_{ik}^{(b+1)})^m}, i = 1,2,\cdots,c \tag{2}$$

step4:如果 $|\boldsymbol{P}^{(b)} - \boldsymbol{P}^{(b+1)}| < \epsilon$,则停止并输出 $\boldsymbol{U}, \boldsymbol{P}$,否则 $b = b + 1$,转向 step2。

根据算法得到的隶属度矩阵 $\boldsymbol{U}$ 和矩阵原型 $\boldsymbol{P}$。得到最终聚类结果 c 种类别。

2 实　验

2.1 KDD Cup 1999 数据集

在实验阶段,选用 KDD Cup 1999 数据集[6],它来源于 DARPA 入侵检测评估计划。该数据一共提供了 4900000 条关于连接的数据记录,对应于提供的每一个 TCP/IP 连接,除了一些基本属性(如协议类型、传送的字节数等)外,还利用领域知识扩展了一些属性(例如登录失败的次数、文件生成操作的数目等),某些属性是在计算过去 2 秒钟之内信息的基础上得到的,例如在过去 2 秒钟连接到同一个服务的连接数目。每个连接共有 41 种定性和定量的特征,其中有 8 个属性是离散型的变量,其余是连续型的数字变量。

入侵数据有 4 大类,24 小类。分别是:DOS(Denial of Service) 攻击,例如 SYN 洪流;U2R(未授权的提升权限)攻击,例如各种缓冲区溢出攻击;R2U(未授权的远程登录) 攻击,例如猜测密码;PROBING 攻击,例如端口扫描。根据 Pal 等[7]的从聚类有效性角度考虑,设置 m 取值为[1.5,2.5],后续试验中 m 取 2.1 效果较好,故本算法令 $m = 2.1$。

2.2 样本的选取及实验结果

以随机方式重新建立 6 个样本集,每个集合包含 1000 个正常实例和 100 个入侵实例。在算法实现过程中忽略类标识属性,其仅供算法结果分析之用。实验样本集结构如表 1 所示。

表 1 实验数据结构

	实例数	正常实例数	入侵实例数	攻击类型数	分类攻击数
数据集 1	1100	1000	100	22	4
数据集 2	1100	1000	100	20	4
数据集 3	1100	1000	100	3	2
数据集 4	1100	1000	100	4	2
数据集 5	1100	1000	100	3	3
数据集 6	1100	1000	100	8	3

表 2 为仿真实验结果,其中包含两个重要检测性能参数:

表 2 实验结果

	正常实例数	入侵实例数	检测率	误检率
数据集 1	999	49	49%	0.1%
数据集 2	1000	41	41%	0%
数据集 3	986	100	100%	1.4%
数据集 4	994	50	50%	0.6%
数据集 5	1000	100	100%	0%
数据集 6	944	20	20%	5.6%
平均值	999.83	60	60%	1.283%

(1)检测率:$D_r = n_i/N_i$ 表示入侵行为的检测比例,其中 n_i 为检测出的入侵实例数目,N_i 为数据集中入侵实例总数。

(2) 误检率:$F_r = (N_i - n_i)/N_n$ 表示误将入侵判断为正常行为的比例,其中 N_n 表示数据集中正常实例数目。

这两项指标能充分反映算法的检测能力。在仿真实验中,入侵实例与正常实例之比为 10∶1 ,而平均检测率仍大于 50%,平均误检率保持约 1.3%。这充分表明算法对于未知攻击检测的可行性和有效性。

3 结 论

文中针对网络入侵的异常检测问题,利用基于模糊 C-均值聚类算法进行入侵检测。由于网络流量样本一般具有混合性属性,因此给出了一种新型的基于属性值范围的加权相似性度量方法。最后,利用 KDD Cup 1999 数据集对该算法进行了实验。实验结果表明,此算法异常入侵检测问题是可行、有效的,具有良好的可扩展性。

参考文献:

[1] 唐正军.网络入侵检测系统的设计与实现[M].北京:电子工业出版社,2002.

[2] 胡昌振.网络入侵检测原理与技术[M].北京:北京理工大学出版社,2006.

[3] Han Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. [s.l.]: Morgan Kaufman, 2001.

[4] 罗 静,董 晟,华 鹏.一种基于克隆的模糊 C-均值入侵检测方法[J].微机发展,2004,14(3):107-109.

[5] 高新波.模糊聚类分析及其应用[M].西安:西安电子科技大学出版社,2004.

[6] KDD99. KDD99 cup dataset[DB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99.

[7] Pal N R, Bezdek J C. On clustering for the fuzzy c-means model[J]. IEEE Trans FS, 1995, 3(3): 370-379.

(上接第 177 页)

W, ed. Advances in Cryptology - EUROCRYPT'91. Berlin: Springer-Verlag, 1991: 257-265.

[2] Shamir A. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612-613.

[3] Desmedt Y, Frankel Y. Threshold cryptosystems[C]//In: Advances in Cryptology - Crypto89, Lectures Notes in Computer Science 435. Berlin : Springer-Verlag, 1989: 307-315.

[4] Desmedt Y, Frankel Y. Shared generation of authenticators and signatures[C]//In : Advances in Crytology. Crypto'91. Berlin: Springer-Verlag, 1991: 457-469.

[5] Wang G. On the security of the Li-Hwang-Lee-Tsai threshold group signature scheme[C]//In: Proceedings of Information Security and Cryptology (ICISC 2002). Berlin: Springer-Verlag, 2003: 75-89.

[6] Camenisch J, Stadler M. Efficient Group Signature Schemes for Large Groups[C]//Advances in Cryptology - CRYPTO'97. [s.l.]: Springer-verlag, 1997: 410-424.

[7] 郭兴阳,张 权,唐朝京.一种动态门限群签名方案的安全性分析[J].国防科技大学学报,2005,27(4): 71-74.