文章编号:1000 - 582X(2005)07 - 0074 - 04

基于模糊聚类理论的入侵检测数据分析*

鲜 继 清[1],郎 风 华[2]
(重庆邮电学院 1. 自动化学院; 2. 计算机学院,重庆 400065)

摘 要:入侵检测系统是网络和信息安全构架的重要组成部分,主要用于区分系统的正常活动和可疑及入侵模式,但是它所面临的挑战是如何有效的检测网络入侵行为以降低误报率和漏报率. 基于已有入侵检测方法的不足提出利用模糊 C - 均值聚类方法对入侵检测数据进行分析,从而发现异常的网络行为模式. 通过对 CUP99 数据集的检测试验表明该方法不但可行而且准确性及效率较高.

关键词:入侵检测;异常检测;模糊聚类;模糊 C - 均值聚类
中图分类号:TP393.08;TP301.6 **文献标识码**:A

入侵检测系统是一种主动的安全防护系统,它提供了对内、外部攻击和误操作的实时保护,并能在网络系统遭遇危害之前拦截和响应入侵,因此它已成为网络安全系统深层防御的重要组成部分.

近年来,国内外学者已提出大量的入侵检测方法,如统计方法[1]、贝叶斯推理方法[2]、机器学习[3]方法,神经网络[4],数据挖掘[5],遗传算法[6],HMM[7]、以及基于 SLT 和 SVM[8]的方法等. 但上述方法存在一个缺点,即对建立检测模型的训练数据要求较高:必须是"干净"的数据并且必须包含检测对象的大多数正常行为,而同时做到这两点是非常困难的. 因此,笔者利用模糊聚类分析中的模糊 C - 均值法(FCM)寻找网络环境下的入侵行为,测试实验表明该方法是检测网络异常数据的有效工具.

1 FCM 聚类的原理

模糊 C - 均值(简称 FCM)算法[9]是最重要也是最流行的模糊聚类算法之一. 1973 年 Dunn 首先提出了 FCM 算法的一个特例,同年 Bezdek 将 Dunn 的算法推广到 $m>1$ 的情形,之后又出现了许多相关的算法和各种间接的推广. 鉴于权重指数 m 等参数在 FCM 聚类算法中的重要作用,文中对其进行了专门讨论.

1.1 模糊 C - 均值法

给定数据集 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 是 s 维模式空间的一个特征向量集,根据某种相似性度量,该集合被聚合成 c 个子集 $X_1,X_2,\cdots,X_c$,$2\leqslant c<n$,这 c 个子集组成特征向量集 X 的一个模糊划分,μ_{ik} 表示特征向量 x_i 属于子集 X_k 的隶属度,从而可得模糊分类

$$U=[\mu_{ik}]_{c\times n}=\begin{bmatrix}\mu_{11} & \mu_{12} & \cdots & \mu_{1n}\\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \mu_{c1} & \mu_{c2} & \cdots & \mu_{cn}\end{bmatrix}_{c\times n},$$

且满足 $0\leqslant\mu_{ik}\leqslant 1(1\leqslant i\leqslant c,1\leqslant k\leqslant n)$;$\sum_{i=1}^{c}\mu_{ik}=1(1\leqslant k\leqslant n)$;$0<\sum_{k=1}^{n}\mu_{ik}<1(1\leqslant i\leqslant c)$. 令所有 $U=[\mu_{ik}]_{c\times n}$ 的集合记 M_{fcn},即 $M_{fcn}=\{U=[\mu_{ik}]_{c\times n}\mid\mu_{ik}\in[0,1],\forall i,k;\sum_{i=1}^{c}X_{ik}=1,\forall k;n\geqslant\sum_{k=1}^{n}X_{ik}\geqslant 0,\forall i\}$,称 M_{fcn} 为 X 的软 c 划分空间. 用矩阵 R^{cs} 表示所有的实 $c\times s$ 阶矩阵集合,令 $v=(v_1,v_2,\cdots,v_c)^T\in V$ 是聚类中心,其中 $v_i\in V$ 是类 $i(1\leqslant i\leqslant c)$ 的聚类中心. 则 FCM 的目标函数可表示为 $J_m(u,v)$:

$$J_m(u,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m\|x_k-v_i\|^2,$$

式中 $\|x_k-v_i\|^2=(x_k-v_i,x_k-v_i)$,$1<m<+\infty$. Dunn 证明了求上述泛函极小值问题是可解的. Bezdek[10]已证明,(u^*,v^*) 是 $J_m(u,v)$ 的局部极小值的必要条件是:

* 收稿日期:2005 - 04 - 20
基金项目:国家自然科学基金网络与信息安全重大研究项目(9030404)
作者简介:鲜继清(1946 -),男,重庆人,重庆邮电学院副教授,主要研究方向:通信系统与网络.

$$v_i^* = \frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m}(i = 1,2,\cdots,c),$$

$$u_{ik}^* = \frac{1}{\sum_{j=1}^{c}(\frac{d_{ik}^*}{d_{jk}^*})^{\frac{1}{m-1}}}(i = 1,2,\cdots,c;k = 1,2,\cdots,n),$$

同时 Bezdek 给出了 $m \geqslant 1$ 且 $x_k \neq v_i$ 的迭代算法.

1.2 模糊 C - 均值算法

FCM 算法的基本思路是用迭代方法求解式 v_i^* 和 u_{ik}^* 式,直至满足某个终止条件,具体步骤如下:

1) 给定聚类数目 c,权指数 m 以及迭代标准 ε;

2) 选定初始的聚类中心 $v = (v_1, v_2, \cdots, v_c)^T$;

3) 用当前的聚类中心计算隶属度函数 u_{ik}^*;

4) 用当前的隶属度函数更新计算各类聚类中心 v_i^*;

5) 若前后两次计算所得的模糊矩阵之间的距离不大于 ε,则算法终止,否则转步骤 3).

用上面算法得到的 (U^*, V^*) 是相对于分类数 c,加权指数 m 以及 ε 等条件下的最优解. 因此,对于不同的 m 值,就会有不同的模糊 C - 划分,故必须要考虑 m 的最优取值问题.

1.3 最优权重指数 m^* 的确定

权重指数 m 是 FCM 算法中最重要的参数之一,它的取值不但影响着目标函数的凹凸性,还控制着聚类的模糊性、模糊类间的分享程度、噪声、目标函数的凹凸性及算法的收敛性等. 理论上讲,m 的取值范围为 $[1, +\infty]$,但是,当 m 接近 1 时,FCM 退化为硬 C - 均值算法;当 m 趋于无穷时,FCM 的唯一解是数据集的质心,失去了划分功能. 同时若 m 取的过小,则 FCM 算法的抗噪性能变差;反之若取的过大,则得不到准确的原型模式. 最优的权重指数应使得模糊聚类的类内加权误差和最小. 同时还要保证聚类间良好的可分性. 由

$$\frac{\partial J_m(u,v)}{\partial m} = \sum_{i=1}^{c}\sum_{k=1}^{n}[(\mu_{ik})^{m-1} \| x_k - v_i \|^2][\mu_{ik}\ln(\mu_{ik})]$$

可知,$J_m(u,v)$ 随 m 的增加而单调递减. 研究表明目标函数 $J_m(u,v)$ 对参数 m 的偏导数存在一个极小点,并且该点恰好在 Bezdek 的经验范围 $[1.1,5]$ 之内,因此得到一种最优的 m 选取方法:

$$m^* = \left\{m \left| \frac{\partial}{\partial m}\left(\frac{\partial J_m(u,v)}{\partial m}\right) = 0\right.\right\}.$$

考虑到入侵检测的实时性,笔者对上式加以简化得:

$$m^* = \arg\left\{\min_{\forall m}\left|\left\{\frac{\partial J_m(u,v)}{\partial m}\right\}\right.\right\}.$$

1.4 相对状态特征值

FCM 算法得到的最优分类矩阵 U^* 是模糊矩阵,对应的分类是软分类. 为了使 U^* 清晰化,一般采用最大隶属度原则,但最大隶属度原则有时不能适用,若对该原则误用,会导致不合理的判断. 因此采用相对状态特征值.

由最优分类矩阵 U^* 可知,样本 x_i 对类别 $l = 1,2,\cdots,c$ 的相对隶属度为 $\mu_{1i}^*, \mu_{2i}^*, \cdots, \mu_{ci}^*$ 为状态变量 l 对应的相对隶属度权重,其总和为 $L(x_i) = \sum_{l=1}^{n}\mu_{li}^* \times l$ 为相对状态特征值,可见 $L(x_i)$ 利用了状态变量 l 的全部相对隶属度信息,因此根据 $L(x_i)$ 判决样本 x_i 归属于何种状态更加确切.

2 仿真试验及结论

为了验证提出的基于模糊 C - 均值聚类的入侵检测数据分析方法的性能,笔者利用 KDD CUP99[11] 数据进行了测试.

2.1 数据预处理

笔者实验中所用的数据集取自美国国防部高级研究计划局 DARPA1999 年入侵检测评估项目,这批数据集是 Wenke Lee 等人在 1998 DARPA 作 IDS 获得的数据基础上恢复出来的连接信息,它约含有 500 万条记录,其中有大量的正常网络流量和各种攻击,具有很强的代表性.

在运用聚类算法之前,需要对数据的属性值进行归一化预处理,主要是因为整个原始的测试数据中包括符号型的和数字型的属性特征变量,同时不同的属性特征有不同的度量标准,若不进行预处理的话就有可能出现较大的偏差. 为了消除各种因素对距离产生的影响,需要对属性值进行预处理. 处理的方法为:1) 对于符号型的属性特征变量值,应将其转换为数字型特征变量值. 2) 将每个数据的每个分量归一化到 $[0, 1]$ 区间上:

$$x_{ik}^* = (x_{ik} - \min(x_{ik}))/(\max(x_{ik}) - \min(x_{ik})),$$

按照上述的方法把所有待检测的数据进行归一化处理.

2.2 最优权重系数 m^* 的仿真

在选定分类数 $c = 2$ 的前提下,$\partial J_m(u,v)/\partial m$ 与 m 的函数图像如图 1 所示.

由 $m^* = \arg\left\{\min_{\forall m}\left|\left\{\frac{\partial J_m(u,v)}{\partial m}\right\}\right.\right\}$. 及图 1 可知,

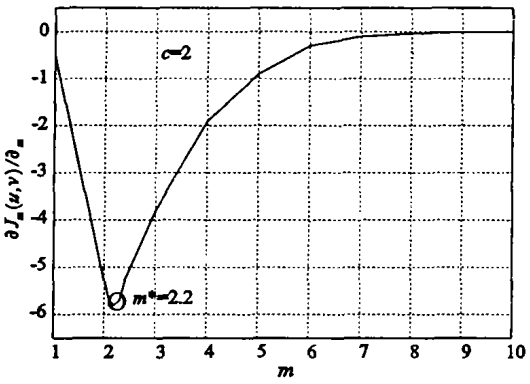

图1 m 与的函数图

$m^*=2.2$ 为最优值,这与 FCM 算法中常取的 $m=2$ 基本一致.

2.3 试验结果

从 KDD CUP99 选 10 组不同检测数据集,每组包括 2 万条检测数据,分别对每组数据进行人侵检测实验,反复实验试探 μ_{ik} 最恰当的取值,并在 $c=1, m=2.2$ 的条件下应用 FCM 算法对其进行聚类. 在此笔者采用以下 2 个参数来描述系统的性能:误报率 (FPR: False Positive Rate) = 正常连接误报为异常连接的数目/正常连接总数目; 检测率(DR:Detection Rate) = 已检测出来的异常连接的数目/异常连接总数目,记录结果如图 2 所示.

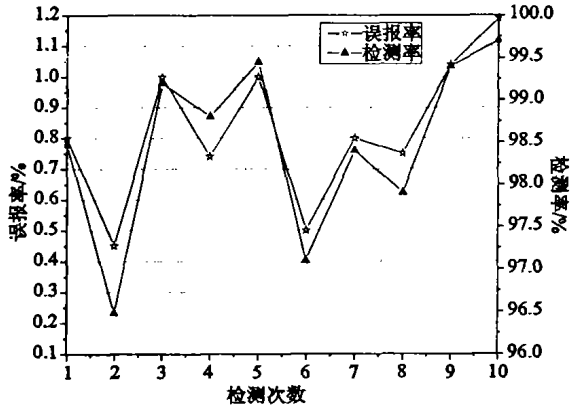

图2 实验次数与误报率及检测率的关系图

检测率和误检率是 IDS 最重要的性能指标,检测率与误检率总是紧密相关,增加检测率常常要以误检率的增加为代价,而误检率偏高使系统对原本不是攻击的事件产生了错误的警报,将导致 IDS 的功效降低. 因此,既能增加检测率又能降低误检率是 IDS 最希望达到的目标. 从图中可以看到误报率控制在 2% 以下时,检测率可达 95% 以上,与其它的检测算法相比性能有极大的提高. 但同时也存在一些问题,比如仍然有某些分布很特殊的数据集不能得到很好的聚类效果,还需进一步改进.

3 结 论

所提出的基于模糊 C - 均值聚类的入侵检测数据分析方法是利用 FCM 算法对网络异常行为进行检测. 通过对 KDD CUP99 数据集的测试实验表明此法可以较好的提高检测效率和降低误报率,因而具有较高的可行性和实用性. 由于 FCM 聚类对初始化数据很敏感,并且最优权重指数的选取也期待更优良的算法,因此进一步的工作可将神经网络、遗传算法等思想与 FCM 聚类思想相结合,以进一步降低入侵检测系统的误报率和漏报率,改善系统的综合性能.

参考文献:

[1] ANDERSON J P. Computer Security Threat Monitoring and Surveillance[R]. James P Anderson Co, Fortwashingon, Pennnsylvania, 1980.

[2] LUNT T F, TAMARU A. A Real Time Intrusion Detection Expert System(IDES)[R]. Compputer Science Laboratory, SRI International, Menlo Park, California, 1992.

[3] WHITE G B, FISCH E A, POOCH U W. Cooperating Security Managers: a Peer Based Intrusion Detection System[J]. IEEE Network, 1996, 10(1): 20 - 23.

[4] ANUP K GHOSH, AARON SCHWARZBARD. A Study in Using Neural Networks for Anomaly and Misuse Detection [Z]. The 8th USENIX Security Symposium, Washington D C, 1999.

[5] HOCHBERG J, JACKSON K, STALTINGS C,et al. NADIR: an Automated System for Detecting Network Intrusion and Misuse [J]. Computer and Security, 1993, 12 (3): 235 - 248.

[6] BALAJINATH B, RAGHAVAN S V. Intrusion Detection Through Learning Behavior Model [J]. Computer Communication, 2001, 24(12): 1 202 - 1 212.

[7] JHA S, TAN K, MAXION R A. Markov Chains, Classifiers and Intrusion Detection[Z]. The 14th IEEE Computer Security Foundations Workshop, Canada, 2001.

[8] BERNHARD SCHOLKOPF, JOHN C PLATTZ. Estimating the Support of a High-dimensional Distribution [J]. Neural Computation, 2001, 13(7): 1 443 - 1 472.

[9] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York: Plenum press, 1981.

[10] BEZDEK J C. Cluster Validity with Fuzzy Sets[J]. J Cybernt, 1974, 3(3): 58 - 72.

[11] KDD99. KDD99 Cup dataset[DB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kd - dcup99.html. 1999.

Fuzzy Clustering Theory for Analyzing Intrusion Detection Data

XIAN Ji-qing[1], *LANG Feng-hua*[2]

(1. School of Automation, Chongqing University of Posts and Telecommunication, Chongqing 400065, China;
2. School of Computer Science and Technology, Chongqing University of Posts & Telecommunication, Chongqing 400065, China)

Abstract: Intrusion detection system is an important component of the computer and information security framework. Its main goal is to differentiate between normal activities of the system and behaviors that can be classified as suspicious or intrusive, and its main challenge is to efficiently detect intrusion detection behaviors for reducing false positive rate and false negative rate. In view of the disadvantages of the existing intrusion detection methods, fuzzy c-means(FCM) clustering method is used to analyze intrusion detection data in order to detect anomaly network behavior patterns. Experimental results on the CUP99 data set data show that this method can not only feasible but also improve the accuracy and efficiency.

Key words: intrusion detection; anomaly detection; fuzzy clustering; fuzzy c-means clustering

(编辑　吕赛英)

(上接第 70 页)

Fast Volume Rendering Technology Based on Programmable GPU

ZHANG Jian-xun, LIU Quan-li, CHEN Zhuang

(School of Computer Science & Engineering, Chongqing Institute of Technology, Chongqing 400050, China)

Abstract: Techniques of programmable vertex shader and pixel shader have been integrated in newly developed graphics hardware armed with powerful Graphics Processing Unit(GPU) in recent years, and as a result, real-time volume rendering can be implement. First, rendering pipeline, hardware architectures on per-pixel shading and fast rendering principium of the modern programmable GPU are explained in detail. Second, technology on how to analyze and solve volume rendering problems is described. Finally, maximum intensity projection (MIP) method rendering medical volume data have been implemented based on programmable Graphics Processing Unit. In a performance test, spent time rendering medical volume data based on programmable pixel shader in GPU is obviously less than spent time rendering it do without programmable pixel shader.

Key words: graphics processing unit; real-time volume rendering; maximum intensity projection; Cg

(编辑　张小强)