

粒子群模糊聚类算法在入侵检测中的研究

李 锋

(广东交通职业技术学院, 广东 广州 510650)

摘 要:目前模糊C均值聚类算法广泛应用于入侵检测算法中,但是存在聚类数目难以确定,目标函数的局部极小点使得算法容易陷入局部最优的现象,影响入侵检测的准确率。鉴于此,文中提出一种基于粒子群算法的模糊聚类算法,引入PSO全局搜索能力和粒子翻转变异操作,避免传统C均值聚类算法对孤立点敏感,容易陷入局部最优,过早收敛的问题。最后通过实验结果表明,新算法检测率明显优于C均值聚类算法,能很好地应用于目前入侵检测系统之中。

关键词:模糊C均值聚类算法;粒子群算法;模糊聚类;入侵检测

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2014)12-0138-04

doi:10.3969/j.issn.1673-629X.2014.12.032

Research on Fuzzy Clustering Algorithm Based on PSO in IDS

LI Feng

(Guangdong Communication Polytechnic, Guangzhou 510650, China)

Abstract: Fuzzy C-means clustering algorithm is widely used in intrusion detection currently. But this algorithm has some shortcomings that is difficult to determine the clustering number and easy to fall into the local minimum when iterating, which can affect the accuracy of intrusion detection system. In view of this, propose a fuzzy clustering algorithm based on PSO algorithm, through introducing the PSO global search ability and particle inverting operation, avoid the problem of falling into local minimum and premature convergence. Finally, the experimental results show that the new algorithm has higher detection rate than the C-mean clustering algorithm, which can be well applied to intrusion detection systems.

Key words: FCM algorithm; PSO algorithm; fuzzy cluster; IDS

0 引言

入侵检测系统(IDS)是一种主动防御体系,它从网络环境中采集分析数据,通过检测引擎判断可疑攻击和异常事件,在系统受到危害之前拦截攻击行为。然而随着网络入侵技术层出不穷,相应的检测技术已明显滞后于攻击技术的更新,如何提高IDS检测效率一直是研究重点。

聚类是把物理或抽象对象的集合按相似度分成多个类的过程,俗称物以类聚。由聚类算法归类的簇是一组具有相似数据对象的集合,而与其他簇中的对象存在较大差异。聚类算法应用于入侵检测系统基于以下两个假设:

(1)网络中正常数据的流量远远大于攻击数据的流量;

(2)攻击数据流量在某些属性取值上偏离正常取值范围^[1-2]。

由此,入侵检测问题可以转化为利用聚类算法查找孤立的问题^[3]。模糊C均值聚类算法是一种基于目标函数的划分算法,目前已广泛应用于异常检测IDS系统之中。

1 模糊C均值聚类算法步骤

对给定数据样本集 $X = \{x_1, x_2, \dots, x_n\}$ 共有 n 个样本点,每个样点包含 s 个属性,构成 s 维空间特征向量集^[4-5]。模糊聚类就是要将样本集 X 划分成 c 个聚类中心。在模糊划分中,每一样本点不能严格归属为某一大类,而是根据一定隶属度划分到某一类。例如将数据集 n 个数据划分为 c 个聚类 X_1, X_2, \dots, X_c , 每个聚

收稿日期:2014-03-07

修回日期:2014-06-10

网络出版时间:2014-10-23

基金项目:2012年广东省高等学校教学质量与教学改革工程省级精品课程(粤教高函[2013]13号);2013年广东省高职教育教学指导委员会教学教改项目(xxjs-2013-2001);2013年广东省高职高专校长联席会议教改项目(GDXLHQ012)

作者简介:李 锋(1981-),男,广东龙川人,硕士,讲师,研究方向为网络安全和图像处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1520.041.html>

类中心为 q_j , 组成 $Q = \{q_1, q_2, \dots, q_c\}^{[6-7]}$ 。令 u_{ij} 表示第 i 个样本点第 j 类的隶属度, 满足以下条件:

$$\begin{cases} u_{ij} \in [0, 1] \\ \sum_{j=1}^c \mu_{ij} = 1 (1 \leq i \leq n) \\ 0 < \sum_{j=1}^m u_j < n (1 \leq j \leq c) \end{cases} \quad (1)$$

模糊 C 均值聚类算法采用误差平方和函数作为聚类目标函数, 公式如下:

$$J_m(u, q) = \sum_{i=1}^n \sum_{j=1}^c (\mu_{ij})^m \|X_i - V_j\|^2 \quad (2)$$

其中, $\|X_i - V_j\|^2$ 为样本点 X_i 与聚类中心 V_j 之间的距离; $m \geq 1$ 是模糊加权参数。算法就是求在 u_{ij} 满足条件时目标函数 J 的最小值。迭代过程中 u 和 v 计算按如下公式:

$$u_{ij} = \begin{cases} \left[\sum_{k=1}^c \frac{\|X_i - V_j\|^{\frac{2}{m-1}}}{\|X_i - V_k\|^{\frac{2}{m-1}}} \right]^{-1}, & \|X_i - V_k\| \neq 0 \\ 1, & \|X_i - V_k\| = 0 (k = j) \\ 0, & \|X_i - V_k\| = 0 (k \neq j) \end{cases} \quad (3)$$

$$\text{其中, } v_j = \frac{\sum_{i=1}^n (u_{ij}^m \cdot x_i)}{\sum_{i=1}^n (u_{ij}^m)}.$$

模糊 C 均值聚类算法迭代过程如下:

- (1) 给定类别数 c , 模糊度 m 和允许误差值 ε ;
- (2) 初始聚类中心 $v_j^L (j = 0, 1, \dots, c)$, L 表示迭代步数;
- (3) 根据式(1)计算隶属度矩阵 U^L ;
- (4) 根据式(3)计算 V^{L+1} ;
- (5) 计算误差 $e = \|V^{L+1} - V^L\|$, 如果 $e < \varepsilon$, 则算法满足迭代终止条件, 输出聚类结果; 否则 $L = L + 1$, 跳回步骤(3)。

模糊 C 均值聚类算法在异常检测中输出结果为一个 $N \times K$ 模糊划分矩阵, 这个矩阵表示每个数据样本对象对于每个聚类的隶属度, 当隶属度满足异常聚类时即判为攻击行为^[8]。

模糊 C 均值聚类算法通过对网络大量数据流的归类检测异常攻击, 在入侵检测中得到了很好的应用, 但同时存在以下几个问题。

(1) 聚类数目需要事先设置匹配参数, 选择合适的聚类数量是正确聚类的前提, 但在实际应用中聚类数目往往难以确定;

(2) 该算法本质上是一种寻优技术, 但是目标函数存在许多局部极小点, 使得算法每次迭代都沿目标函数减小的方向收敛, 陷入局部最优现象, 影响入侵检

测准确率。

2 基于粒子群算法模糊聚类

2.1 标准粒子群算法

粒子群算法 (PSO) 是基于鸟类群体行为研究的模拟算法^[9-10]。鸟群在封闭空间随机搜索食物, 并且在这个空间只有一个全局最优值。假如所有鸟都只知道当前位置与搜索食物之间距离, 那么找到全局最优解的最优方案就是从身边最近的鸟周围区域进行搜寻。在粒子群算法中, 寻找最优问题的每个解对应搜索空间的每只鸟, 称为粒子。每个粒子的初始化向量代表鸟的飞行位置和速度, 每个粒子通过寻找附近粒子迭代搜寻最优解, 具体算法如下:

假设在一个 D 维搜索空间中, 有 N 个粒子组成的粒子群 $X = (X_1, X_2, \dots, X_N)^T$, 其中第 i 个粒子位置为 $X_i = (X_{i1}, X_{i2}, \dots, X_{iD})^T$, 速度为 $V_i = (V_{i1}, V_{i2}, \dots, V_{iD})^T$, 第 i 个粒子极值为 $P_i = (P_{i1}, P_{i2}, \dots, P_{iD})^T$, 种群全局极值为 $P_g = (P_{g1}, P_{g2}, \dots, P_{gD})^T$, 每个粒子找到下一粒子后按以下公式更新当前位置和速度^[11-12]:

$$v_{id}^{k+1} = w v_{id}^k + c_1 \text{rand}_1^k (p_{id}^k - x_{id}^k) + c_2 \text{rand}_2^k (p_{gd}^k - x_{gd}^k) \quad (4)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} \quad (5)$$

$$w(k) = w_{\max} - k * \frac{(w_{\max} - w_{\min})}{k_{\max}} \quad (6)$$

式中, k 表示迭代次数; c_1 和 c_2 为加速系数; rand 是 $[0, 1]$ 区间选取的随机数; p_{id}^k 是第 i 个粒子的个体极值在第 d 维的分量; p_{gd}^k 是群体全局极值在第 d 维的分量; x_{id}^k 和 v_{id}^k 分别是第 i 个粒子经 k 次迭代后的第 d 维位置和速度; w 是粒子保持运动惯性权重。

粒子通过不断更新当前位置和速度最终找到全局最优解, 完成搜索过程。

标准 PSO 算法采用群体解合作机制迭代产生最优解, 这种基于全局的搜索能力能避免算法陷入局部最优。但是受于种群规模限制, 标准 PSO 算法经过迭代进化后较好个体会逐渐充斥整个种群, 因此找到的解只是全局近优解^[13-14]。为解决这个问题, 文中提出在 PSO 算法基础上增加变异操作, 结合模糊聚类算法用于 IDS 异常检测之中。

2.2 新算法变异操作

新算法在标准 PSO 基础上增加变异操作, 在粒子各维编码中随机选取若干位进行翻转操作。例如将闭区间内 $[A, B]$ 的实数值按照求解精度转变成二进制字符串, 设求解精确到 x 位, 则把 $[A, B]$ 区间分为 $[B - A] * 10^x$ 等份, 再随机产生一个 $[0, d/10 + 1]$ 之间的整数 ξ , 随机选取二进制串中的 ξ 位对其翻转。以二进制串

$(a_{d-1}a_{d-2}\cdots a_0)$ 为例转化为区间 $[A,B]$ 内对应实数。第一步将二进制串 $(a_{d-1}a_{d-2}\cdots a_0)$ 转化为十进制数：

$$(a_{d-1}a_{d-2}\cdots a_0) = \sum_{i=0}^{d-1} (a_i * 2^i) = x' \tag{7}$$

第二步 x' 对应 $[A,B]$ 内的实数 x 为：

$$x = A + x' * \frac{B - A}{2^d - 1} \tag{8}$$

2.3 新算法适应度函数的确定

在粒子群模糊聚类中,模糊聚类迭代过程转换为粒子群算法的种群进化过程,粒子群中的每一粒子代表某一聚类中心的选取。因此,每个粒子适应度大小代表聚类中心聚类效果的优劣。据此,文中利用误差平方和函数作为新算法适应度函数,见式(9)。

$$\text{fitness} = \frac{1}{\sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \cdot d_{ij}^2} \tag{9}$$

2.4 新算法流程

设聚类样本集合为 $X(X = (X_1, X_2, \cdots, X_m))$, X 是 n 维向量,需把样本空间聚为 k 类。由于 PSO 算法中每个粒子代表一个全局可行解,所以新算法的每个粒子就是聚类中一个簇中心集合 $X_i = \{C_{i1}, C_{i2}, \cdots, C_{ik}\}$,其中 $C_{ij}(j=1,2,\cdots,k)$ 是 n 维向量,表示第 i 个粒子第 j 类中心点的坐标。

新算法流程如下：

- (1) 初始化,给定聚类数目、加速因子 c_1, c_2 , 群体规模 n , 惯性权重最大值和最小值分别为 w_{\max} 和 w_{\min} , 粒子群最大迭代次数为 t_{\max} 。
- (2) 初始化聚类中心 Q' 并对所有数据属性对象编码,构成 n 个第 1 代粒子。每个粒子当前位置为 P_{id} , 当前粒子群中最优解的粒子位置是 P_{gd} 。
- (3) 根据式(3)计算隶属度 u_{ij} 。
- (4) 根据式(9)计算各粒子适应度值,比较前一次迭代计算的最优位置目标函数值与本次目标函数值。如果当前值更优,则 $X_{id}=P_{id}$, 否则保持原 P_{id} 值。
- (5) 比较每个粒子适应度值和群体共同计算的最优位置适应度值,如果更优,则将其作为群最优。
- (6) 根据式(4)和式(5)调整当前粒子的速度和位置。
- (7) 随机选取若干粒子根据式(7)和式(8)进行变异操作。

(8) 如果达到结束条件,如误差小于给定值或超过最大迭代次数,则算法结束,输出聚类结果;否则跳回步骤(2)。

3 实验分析

3.1 聚类算法仿真实验

仿真实验采用 Iris 数据集,一共分为 3 类,每类含

50 个样本,共 150 个样本,包含萼片宽度、萼片长度、花瓣宽度和花瓣长度四种属性。最大种群寻优次数为 150 代,种群大小为 50,两种算法仿真结果使用 Matlab 绘制,见图 1 和图 2。从图中可以看出,新算法聚类正确率明显高于模糊 C 均值聚类算法。

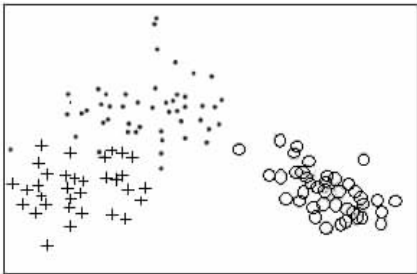


图 1 模糊 C 均值聚类算法

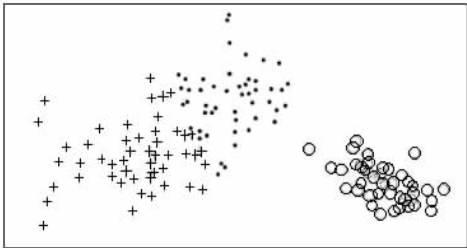


图 2 新算法

3.2 异常检测实验结果

为评价新算法在异常检测中的检测准确率,文中采用 KDDCUP99 数据集进行测试,抽样选取 2 000 条记录作为测试样本集。聚类数目 c 取值为 2,模糊系数 m 取值为 4,并将聚类结果划分为正常数据和异常数据。表 1 和表 2 分别是两种算法在 IDS 入侵检测中的实验结果。

表 1 模糊 C 均值聚类算法结果

数据集	正常实例	攻击实例	检测率	错检率	目标函数值
1	2 000	200	0.835	0.027 5	216.85
2	2 000	200	0.817	0.023 9	231.64
3	2 000	200	0.786	0.031 5	257.17
4	2 000	200	0.881	0.021 4	248.26
5	2 000	200	0.912	0.045 1	268.64
平均值	2 000	200	0.846	0.029 9	244.51

表 2 新算法结果

数据集	正常实例	攻击实例	检测率	错检率	目标函数值
1	2 000	200	0.877	0.031 2	315.78
2	2 000	200	0.836	0.021 4	289.25
3	2 000	200	0.821	0.025 6	352.74
4	2 000	200	0.945	0.014 6	289.36
5	2 000	200	0.845	0.035 2	277.29
平均值	2 000	200	0.865	0.025 6	304.88

从实验结果可以看出,新算法无论是检测率还是误检率都明显优于传统模糊 C 均值聚类算法。这是由于新算法引入 PSO 全局搜索功能和变异操作,避免

C 均值聚类算法对孤立点敏感,容易陷入局部最优的问题,从而提高入侵检测准确率。

3.3 收敛性分析

两种算法收敛性分析见图 3。从图中可以看出,传统模糊 C 均值聚类算法收敛速度过快,在 200 次左右算法收敛,而检测率并无明显增加,说明算法已陷入局部最优现象。新算法由于增加粒子群全局搜索和变异操作,经 300 次迭代收敛速度才渐渐变缓,趋于收敛,并且检测率明显优于 C 均值聚类算法。

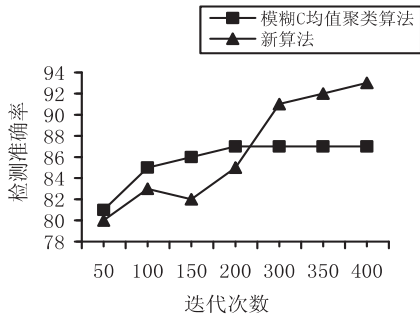


图 3 收敛性分析图

4 结束语

文中提出一种基于粒子群算法的模糊聚类算法,引入 PSO 全局搜索能力和变异操作避免传统 C 均值聚类算法中过早收敛的问题,提高入侵检测系统的检测率。

参考文献:

[1] 李昆仑,黄厚宽,田盛丰,等. 模糊多类支持向量机及其在

[J]. 通信学报,2003,24(2):51-57.

[2] Ye R. A novel chaos-based image encryption scheme with an efficient permutation-diffusion mechanism [J]. Optics Communications,2011,284(22):5290-5298.

[3] 龙卓珉,俞 斌. 针对超混沌系统图像加密算法的选择明文攻击[J]. 计算机工程,2012,38(17):148-151.

[4] 林 冰,蒋国平. 一种基于块置乱和反馈密钥的图像加密算法[J]. 计算机技术与发展,2012,22(5):123-126.

[5] 薛香莲. 一种新的基于超混沌映射的彩色图像加密算法[J]. 计算机应用与软件,2013,30(8):318-321.

[6] Gao Tiegang, Chen Zengqiang. A new image encryption algorithm based on hyper-chaos[J]. Physics Letters A,2008,372(4):394-400.

[7] Zhang G, Liu Q. A novel image encryption method based on total shuffling scheme[J]. Optics Communications,2011,284(12):2775-2780.

[8] 王 静,蒋国平. 一种超混沌图像加密算法的安全性分析及其改进[J]. 物理学报,2011,60(6):83-93.

入侵检测中的应用[J]. 计算机学报,2005,28(2):274-280.

[2] 唐少先,蔡文君. 基于无监督聚类混合遗传算法的入侵检测方法[J]. 计算机应用,2008,28(2):409-411.

[3] 洪飞龙,范俊波,贺 达. 数据挖掘在入侵检测系统中的应用研究[J]. 计算机应用,2004,24(12):82-83.

[4] 杨德刚. 基于模糊 C 均值聚类的网络入侵检测算法[J]. 计算机科学,2005,32(1):86-87.

[5] 王 勇. 模糊 C-均值算法在入侵检测系统中的应用研究[D]. 哈尔滨:哈尔滨理工大学,2007.

[6] 肖 建,白裔峰,于 龙. 模糊系统结构辨识综述[J]. 西南交通大学学报,2006,41(2):135-142.

[7] 肖立中,邵志清,马汉华,等. 网络入侵检测中的自动决定聚类数算法[J]. 软件学报,2008,19(8):2140-2148.

[8] 刘坤朋,罗 可. 改进的模糊 C 均值聚类算法[J]. 计算机工程与应用,2009,45(21):97-98.

[9] 张 敏,于 剑. 基于划分的模糊聚类算法[J]. 软件学报,2004,15(6):858-868.

[10] 刘文远,王颖洁,邓成玉,等. 基于遗传算法的模糊聚类分析[J]. 计算机工程,2004,30(19):117-118.

[11] 张曙红,孙建勋,诸克军. 基于遗传优化的采样模糊 C 均值聚类算法[J]. 系统工程理论与实践,2004,24(5):121-125.

[12] 刘向东,沙秋夫,刘勇奎,等. 基于粒子群优化算法的聚类分析[J]. 计算机工程,2006,32(6):201-202.

[13] Farnstrom F, Lewis J, Elkan C. Scalability for clustering algorithms revisited [C]//Proc of ACM SIGKDD. [s. l.]: [s. n.],2000.

[14] George K, Han Eui-Hong. Hierarchical clustering using dynamic modeling[J]. Computer,1999,32(8):68-75.

[9] Zhu Congxu, Liao Chunlong, Deng Xiaoheng. Breaking and improving an image encryption scheme based on total shuffling scheme[J]. Nonlinear Dynamics,2013,71(1-2):25-34.

[10] 曹光辉,胡 凯,佟 维. 基于 Logistic 均匀分布图像置乱方法[J]. 物理学报,2011,60(11):125-132.

[11] 范九伦,张雪锋. 分段 Logistic 混沌映射及其性能分析[J]. 电子学报,2009,37(4):720-725.

[12] Alvarez G, Li Shujun. Some basic cryptographic requirements for chaos-based cryptosystems[J]. International Journal of Bifurcation and Chaos,2006,16(8):2129-2151.

[13] Amigo J M, Kocarev L, Szczepanski J. Theory and practice of chaotic cryptography[J]. Physics Letters A,2007,366(3):211-216.

[14] Chen Guanrong, Mao Yaobin, Chui C K. A symmetric image encryption scheme based on 3D chaotic cat maps[J]. Chaos, Solitons & Fractals,2004,21(3):749-761.

[15] Wu Y, Noonan J P, Agaian S. NPCR and UACI randomness tests for image encryption[J]. Journal of Selected Areas in Telecommunications,2011,2(4):31-38.

粒子群模糊聚类算法在入侵检测中的研究

作者: [李锋, LI Feng](#)
作者单位: [广东交通职业技术学院, 广东 广州, 510650](#)
刊名: [计算机技术与发展](#) 
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2014(12)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjz201412032.aspx