基于遗传算法的改进模糊C均值算法在入侵检测中的应用

孙秀娟

黑龙江科技学院理学院,哈尔滨 150027

摘要 为克服模糊C均值(FCM)算法对初始化极为敏感且容易陷入局部最优的缺点,将遗传算法和改进的模糊C均值聚类算法相结合,并以检测率和误检测率作为入侵检测算法性能评价的指标,对FCM、改进的FCM、基于遗传的改进FCM 3种聚类算法的入侵检测性能进行仿真分析。仿真实验表明,结合遗传和FCM两种算法的混合算法能够实现优势互补。由于该算法结合了遗传算法,使整个算法的复杂度增加。从入侵检测看,通过增加处理时间而提高了入侵检测率。

关键词 模糊C均值算法;入侵检测;遗传算法;模糊聚类

中图分类号 TP309 **文献标识码** A **文章编号** 1000-7857(2010)15-0049-04

IDS Method Based on Genetic Improved Fuzzy C-Means

SUN Xiujuan

College of Science, Heilongjiang Institute of Science & Technology, Harbin 150027, China

Abstract In order to overcome the shortcomings of the algorithm of Fuzzy C-Means (FCM), namely, the extreme sensitivity to the initial data and the liability to come to local optimum points, this paper proposes a combination of the genetic algorithms and the improved fuzzy C average value methods and the use of the examination rate and the mistaken examination rate as the evaluation indices for the invasion examination algorithm performance. The simulation analysis is carried out for the invasion examination performance of three kinds of algorithms related to FCM, improved FCM and improved FCM based on the heredity. The result indicates that the method, based on the genetic improved FCM algorithm (GIFCM) enjoys higher examination rate and lower mistaken examination rate. This algorithm can find an effective application in exceptionally invasion examination.

Keywords Fuzzy C-Means; ID; genetic algorithm; fuzzy clustering

0 引言

随网络技术及电子商务、电子政务、网络银行等网络业务的快速发展,计算机网络已成为日常生活中的一部分,而日益严重的网络安全问题是制约网络技术发展的一大障碍[1]。目前,保障网络安全的主要技术手段有防火墙、安全路由器、身份认证系统等,这些安全产品大多数属于静态安全技术的范畴。静态安全技术对防止系统非法入侵起到了一定的作用,但从安全管理角度看,仅有防御是不够的,还应采取动态策略[2]。入侵检测(Intrusion Detection,ID)技术能对网络安全实施实时监控、攻击与反攻击等动态保护[3]。因此,入侵检测技术的研究具有很强的现实性和紧迫性。

入侵检测的有效性、可适应性、可扩展性是评价入侵检测系统质量的重要指标。当前的入侵检测系统,通常采用统计分析方法对已知入侵方法和系统脆弱性进行分析,且针对的是具体系统环境和检测方法,使系统的有效性较差,在可扩展性、自适应性方面也很有限[4-5]。入侵检测作为一种主动防御的安全技术,已成为网络安全技术的一个重要手段,并成为当前的热点研究领域。聚类分析是数据挖掘技术中的关键技术,但传统的C均值聚类算法对入侵检测数据进行处理有很多不尽人意的地方,如该聚类算法是局部寻优算法,

收稿日期:2010-01-20;修回日期:2010-06-30

基金项目:黑龙江省教育厅科学技术项目(11551439)

作者简介:孙秀娟,讲师,研究方向为数论与数理统计,电子信箱:ruixuan992006@126.com

聚类的结果对数据输入顺序比较敏感等[6-7]。为克服当前入侵检测系统的缺陷,本研究以数据为中心,把入侵检测当作数据分析过程,即将模糊 C 均值算法应用于入侵检测系统。

1 模糊 C 均值算法

1.1 模糊 C 均值算法原理

模糊 C 均值算法(Fuzzy C-Means,FCM)[8-10]是普通 C 均值算法的改进,可表示每个数据属于各个类别的程度,它通过迭代来优化目标函数,并求取目标函数的极值点,使聚类质量最优。

假设给定数据集 $X=\{x_1, x_2, \cdots, x_n\}$,模糊 C 均值算法把 n 个数据集对象划分为 c 个子类,并给每一个子类定义一个聚类中心,然后根据数据集中的每个数据对象与聚类中心的距离[11],形成一些具有相同性质的模糊子集,每一个数据对象与聚类中心有一个隶属度,并使非相似性指标的目标函数达到最小。FCM 采用模糊划分的方法,使每个给定数据对象用[0,1]间的隶属度来确定其属于各个子类的程度。用 μ_{ik} 表示数据集中对象 x_k 在类别 i 中的隶属度。FCM 具有如下性质:

$$\begin{cases} \mu_{ik}\in[0,\ 1] \\ \sum_{i=1}^{c}\mu_{ik}=1 & \forall k \\ 0<\sum_{k=1}^{n}\mu_{ik}<n & \forall i \end{cases} \tag{1}$$

由此形成一个 $c\times n$ 的模糊矩阵 $U=\{\mu_{ik}\}$。设第 i 个类的聚类中心向量为 v_i,则定义目标函数为

$$J(U,\ V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2 \tag{2}$$

式中,$m>1$,为模糊指数;d_{ik} 表示通常的欧氏距离,$d_{ik}=\|x_k-v_i\|$。$J(U,\ V)$表示各类中样本到聚类中心的加权距离平方和。

聚类准则是在约束 $\sum_{i=1}^{c}\mu_{ik}=1$ 下,选择合适的隶属度 μ_{ik} 和聚类中心向量 v_i,使目标函数 $J(U,\ V)$对 U 的偏微分,通过拉格朗日乘数法,并根据式(1)得

$$\mu_{ik}=\left[\sum_{j=1}^{c}(d_{ik}/d_{jk})^{\frac{2}{m-1}}\right]^{-1} \qquad I_k=\Phi \tag{3}$$

式中,$I_k=\{i|1\leqslant i\leqslant c,\ d_{ik}=0\}$。

特别地,对于 $I_k\neq\Phi$,μ_{ik} 为

$$\mu_{ik}=\begin{cases}0 & \forall i\in\tilde{I}_k \\ 1 & \forall i\in I_k\neq\Phi\end{cases} \tag{4}$$

式中,$\tilde{I}_k=\{1, 2, \cdots, c\}-I_k$。

令 $J(U,\ V)$对 v_i 的偏导为 0,可得到聚类中心向量

$$v_i=\frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m} \qquad i=1, 2, \cdots, c \tag{5}$$

参数 m 是控制算法的标量,用来控制分类矩阵 U 的模糊程度,m 越大标量越模糊。如果 $m=1$,则 FCM 算法退化为 HCM 聚类算法(Hard C-Means)。FCM 聚类需要进行多次迭代计算,以使目标函数取得最小值。因此,若聚类类别数 c 和模糊指数 m 已知,利用式(1)、式(2)即可对数据集 X 进行聚类分析。

1.2 模糊 C 均值算法的应用

FCM 算法需要 2 个参数:聚类数目 c、模糊指数 m。一般情况下,c 应远远小于聚类样本的总个数,同时要保证 c 大于 1。对于 m,它是一个控制算法的柔性参数,如果 m 过大,则聚类的模糊程度越大,聚类效果就越差;如果 m 过小,则算法会接近 HCM 聚类算法。聚类有效性研究表明,m 的最佳取值区间为[1.5, 2.5][12],在无特殊要求下,可取区间中值 $m=2$。算法的输出是 c 个聚类中心点向量和 $c\times n$ 的一个模糊划分矩阵,这个矩阵表示每个数据对象属于每个类的隶属度。根据划分矩阵,按照模糊集合中的最大隶属原则,就能够确定每个数据对象归为哪个类。聚类中心表示的是每个类的平均特征,可以认为是该类的代表点。

基于 FCM 的入侵检测算法过程如下:

1) 按照初始聚类过程所求的聚类数目设定聚类类别数 c,用调整后的聚类中心初始化 c 个聚类中心向量 $v_i(2\leqslant i\leqslant c)$;给定模糊指数 m,设定一个终止迭代误差值 ε,迭代次数 $p=0$。

2) 计算各个数据到聚类中心的距离 d_{ik},如果 $I_k=\Phi$,按照式(3)计算隶属度 $\mu_{ik}(1\leqslant i\leqslant c,\ 1\leqslant k\leqslant n)$,否则,按式(4)计算隶属度。

3) 计算目标函数 $J^{(p)}$,并按照式(5)修正所有聚类中心向量 v_i。

4) 如果 $|J^{(p)}-J^{(p-1)}|\leqslant\varepsilon$,表示收敛,则迭代结束;否则,$p=p+1$,转向过程(2)。

2 FCM 的改进

为了降低孤立点对检测结果的影响,提出一种改进的 FCM。该算法通过给模糊隶属度一个加权值,来减少孤立点对聚类中心的影响,进而达到改善聚类分析结果、提高入侵检测率、降低误检率的目的。

由式(3)可知,普通 FCM 算法的数据集中,每个数据对象的隶属度是基于数据对象到聚类中心的距离的。距离越大,相应的隶属度越小;而聚类中心的更新通过式(1)~式(5)来实现,可看到数据对象的隶属度对聚类中心影响很大。由

$$v_{ik}=\frac{\sum_{k=1}^{n}(\mu_{ik})^m x_{ik}}{\sum_{k=1}^{n}(\mu_{ik})^m} \tag{6}$$

给 μ_{ik} 一个加权值:

$$a_{ik}=\mu_{ik}-\frac{(1-\mu_{ik})}{2}\cdot\mu_{ik} \tag{7}$$

则修改后的聚类中心公式为

$$v_{ik}=\frac{\sum_{k=1}^{n}(a_{ik})^m x_{ik}}{\sum_{k=1}^{n}(a_{ik})^m} \tag{8}$$

由式(7)可见，a_{ik} 与 μ_{ik} 相比，在原有基础上减去了 $\mu_{ik}(1-\mu_{ik})/2$，$(1-\mu_{ik})/2$ 是为了使普通 FCM 算法中的隶属度为 1 时，新加权后的隶属度依然为 1。而 μ_{ik} 则是为了保证在原来的隶属度为 0 时，加权后依然为 0。通过式(7)的修正，可看到处于(0, 1)内的隶属度在经过改进之后，比原来的值有了一定的减少，并且隶属度越小，相应减少得越明显。应用于式(6)，可明显看到隶属度小的数据对象对聚类中心的影响降低。

修改后的 FCM 入侵检测算法，在原有普通 FCM 算法在 ID 中实现的基础上，对第三步求新的聚类中心，改为按照式(8)修改聚类中心，保持其他各个步骤不变，这样可在一定程度上消除孤立点对聚类中心的影响，使求得的聚类中心更加接近最佳点。

3 基于 GIFCM 的算法

3.1 基于 GIFCM 算法的实现

改进模糊 FCM 算法本质上还是 FCM 算法。在执行过程中，$J(U, V)$通过 U 与 V 的迭代，沿着一子序列逐渐收敛到初始 $V(0)$附近的极值点或鞍点[13]。由于 $J(U, V)$是一个多峰的复杂函数，因此，$V(0)$的选择尤为重要。如果 FCM 算法寻优采取对不同的初始 $V(0)$ 执行多次该运算，然后选择其中最好的结果，就相当于在解空间中进行群体搜索。遗传算法[14]是在解空间进行群体搜索，通过遗传操作，群体中的个体得到迭代优化，并逐步逼近最优解。遗传算法的这种全局优化特性，可不断“发现”新的更有希望的搜索区域。但遗传算法的局部搜索能力不如启发式算法，而 FCM 算法的每步迭代都沿着使 J 减小方向进行，有很强的局部搜索能力。基于遗传的改进 FCM 算法结合了这 2 种算法的优点。其基本思想是：用 FCM 算法使群体中每个个体快速趋向各自的极值点，通过遗传算子摆脱个体陷入局部最优的状况，重复进行这样的搜索，直至找到最优解。

由于 FCM 算法具有很好的局部搜索能力，因此，利用该特点，在每次进化迭代后，通过改进的 FCM 将整个数据集聚类，形成新的聚类中心作为下一代的初始聚类中心。当 $|\bar{J}(t)-\bar{J}(t-1)|<\varepsilon$($\varepsilon$ 为小正整数)算法终止，其中 $\bar{J}(t)$ 定义为 $\bar{J}(t)=\frac{1}{n}\sum_{i=1}^{n}J(U_i(t),V_i(t))$，算法流程如图 1 所示。

3.2 基于 GIFCM 在 ID 中的应用

利用 UNIX/Linux 下的 TCPDump 或基于 Windows 平台的 WinPcap 抓包工具，采集来自网络中的各个数据包，它们能够监听和接受网络中所有正在传输的数据包，并把它们记录到文件中。采集到的网络数据包用于后面的异常入侵检测的分析。TCP/IP 的数据连接记录不能直接用于聚类运算。这里采用对数据标准化，再将数据正规化的数据预处理操作。

处理数据集后，对网络数据进行聚类分析，是入侵检测中最重要的环节。使用基于遗传的改进 FCM 算法按照前面介绍的步骤对数据集合进行聚类操作，通过聚类分析，数据集合被划分为大小不一的多个数据集合，由于一般网络数据包中入侵活动只占整个网络活动的很小一部分，因此根据数据集的大小，粗略地将数据集合划分为正常和入侵两部分。

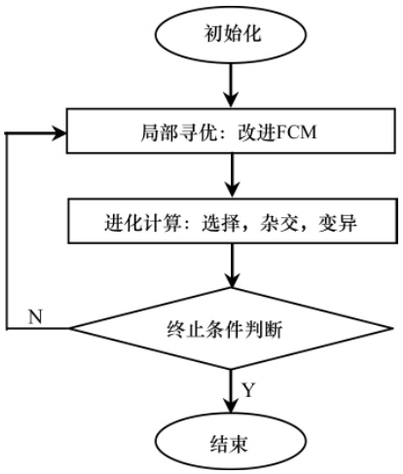

图 1 基于 GIFCM 算法的流程

Fig. 1 Flowchart of GIFCM algorithm

4 实验仿真和结果分析

为评价基于遗传的改进 FCM 聚类的入侵检测算法，在 Intel Pentium 2.4GHz CPU、512MB 内存、Windows XP 操作系统、Matlab 7.0.1 语言编程环境中，以检测率和误检测率作为入侵检测算法性能评价的指标，进行仿真实验。3 种聚类算法入侵检测性能对比结果如表 1 所示。

由表 1 可见，基于遗传的改进模糊 C 均值算法的时间复杂度为 $\sigma(n^3)$，而模糊 C 均值算法和改进的模糊 C 均值算法的时间复杂度均为 $\sigma(n^2)$。即结合了遗传算法后入侵检测的检测性能有了提高，但它以牺牲检测时间为代价。从入侵检测本身的特点来说，这种牺牲还是值得的。因而，基于遗传的改进模糊 C 均值算法应用于异常入侵检测中可行而且有效。

表 1 3 种聚类算法入侵检测性能对比

Table 1 Comparison of three kinds of cluster algorithms in ID

算法	检测率/%	误检测率/%	时间复杂度
FCM	57.6	3.475	$\sigma(n^2)$
改进 FCM	64.23	1.75	$\sigma(n^2)$
遗传+改进 FCM	69.65	1.475	$\sigma(n^3)$

5 结论

将遗传算法引入改进的模糊 C 均值算法中，利用遗传算法全局搜索能力强及受初始化影响小的特点，弥补了模糊 C 均值算法的不足。仿真实验表明，结合遗传和 FCM 两种算法的混合算法能够实现优势互补。由于算法结合了遗传算法，使整个算法的复杂度增加，从入侵检测观点看，通过增加处理时间来提高入侵检测率是值得的。后续研究中，应着重解

决降低算法处理时间的问题,达到更为理想的入侵检测率。

参考文献(References)

[1] 陈兴华. 企业网络信息安全与对策研究 [J]. 农业网络信息, 2009(1): 103-105.
Chen Xinghua. *Agriculture Network Information*, 2009(1): 103-105.

[2] 王海军. 网络信息安全管理体系研究[J]. 信息网络安全, 2008(3): 198-200.
Wang Haijun. *Information Network Security*, 2008(3): 198-200.

[3] 蒋建春, 马恒太, 任党恩, 等. 网络安全入侵检测: 研究综述[J]. 软件学报, 2000, 11(11): 1460-1466.
Jiang Jianchun, Ma Hengda, Ren Dangen, *et al*. *Journal of Software*, 2000, 11(11): 1460-1466.

[4] 杨武, 云晓春, 李建华. 一种基于强化规则学习的高效入侵检测方法[J]. 计算机研究与发展, 2006, 43(7): 1252-1259.
Yang Wu, Yun Xiaochun, Li Jianhua. *Journal of Computer Research and Development*, 2006, 43(7): 1252-1259.

[5] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798-803.
Yao Xian, Dong Chunxi, Yang Shaoquan. *Journal of Software*, 2003, 14(4): 798-803.

[6] 林琳, 王树勋. 基于遗传-模糊聚类的说话人识别方法及其仿真研究[J]. 系统仿真学报, 2006(8): 212-215.
Lin Lin, Wang Shuxun. *Journal of System Simulation*, 2006(8): 212-215.

[7] 陈金山, 韦岗. 遗传+模糊C-均值混合聚类算法 [J]. 电子与信息学报, 2002(2): 88-90.
Chen Jinshan, Wei Gang. *Journal of Electronics & Information Technology*, 2002(2): 88-90.

[8] Ruspini E H. A new approach to clustering [J]. *Inform Control*, 1969, 15(1): 22-32.

[9] Bezedek J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1987.

[10] Dave R N, Bhaswana K. Adaptive fuzzy C-shells clustering and detection of ellipses[J]. *IEEE Transactions on Neural Networks*, 1992(3): 643-662.

[11] 肖位枢. 模糊数学基础及其应用[M]. 北京: 航空工业出版社, 1992.
Wiao Weishu. Fuzzy Mathematics and Its Applications [M]. Beijing: Aviation Industry Press, 1992.

[12] 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, 2004.
Tang Zhengjun. Network intrusion detection system design and implementation [M]. Beijing: Publishing House of Electronics Industry, 2004.

[13] Bezdek J C, Hathaway R, Sabin M, *et al*. Convergence theory for fuzzy C-means, counterexample and repairs [J]. *IEEE Trans on Systems, Man and Cybernetics*, 1987, 17(5): 873-877.

[14] 潘正君, 康立山, 陈毓屏. 演化计算[M]. 北京: 清华大学出版社, 1998: 1-43.
Pan Zhengjun, Tang Lishan, Chen Liuping. Evolutionary Computation [M]. Beijing: Tsinghua University Press, 1998: 1-43.

(责任编辑 陈广仁)