广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(4):229~231
Vol. 23,No. 4 November 2007

一种面向混合数据的自反馈模糊聚类分析算法*
A Feedback Fuzzy Clustering Method Oriented Mixed Data

令狐大智[1],李陶深[2]
LINGHU Da-zhi[1],LI Tao-shen[2]

(1. 广西财经学院计算机与信息管理系,广西南宁 530003;2. 广西大学计算机与电子信息学院,广西南宁 530004)
(1. Department of Computer and Information Management, Guangxi Finance and Economics College, Nanning, Guangxi, 530003, China; 2. School of Computer, Electronics and Information, Guangxi University, Nanning, Guangxi, 530004, China)

摘要:利用动态自反馈理论改造模糊聚类过程,分析入侵数据类型及其在入侵中所起的作用,提出一种面向混合数据的自反馈模糊聚类分析算法,并用算法对KDD99数据集进行对比测试。测试结果显示,本方法能够有效提高入侵检测引擎的检测率,降低其误报率,增强计算机系统抵御入侵及自身免疫能力。

关键词:入侵检测 模糊聚类 自反馈 FCMBP
中图法分类号:TP301.6 文献标识码:A 文章编号:1002-7378(2007)04-0229-03

Abstract: This paper uses dynamic feedback theory to reconstruct fuzzy clustering process and analyze intrusion data types and it's roles in intrusion activity. A feedback fuzzy clustering algorithm oriented mixed data is proposed, and the algorithm's performance is tested by using KDD99 date set. The experimental results show that the method can effectively increase the detection rates of intrusion detection engine and reduce their error rate, and enhance computer systems against intrusion and self-immunity.

Key words: intrusion detection, fuzzy cluster, self-feedback, FCMBP

入侵检测系统是一种主动安全防护系统,它的目的是实现对来自系统内部、外部的攻击及误操作的实时保护。目前,它已成为信息安全层次化综合防御系统的重要组成部分。

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、贝叶斯推理方法、神经网络、数据挖掘、遗传算法、HMM、基于SLT和SVM的方法以及基于模糊逻辑的各种组合方法[1,2]。其中基于模糊逻辑的组合方法进一步地减少了系统的漏报率,增强了防范能力,但它们并没有充分考虑具体方法在模糊聚类过程中,等价性,数据各个字段属性差异及其对产生入侵行为的作用度。虽然漏报率有所降低,但是误报率却居高不下[3]。本文利用动态自反馈理论改造模糊聚类过程,分析入侵数据类型及其在入侵中所起作用,建立面向混合数据的自反馈模糊聚类方法,并在此基础上构建入侵检测模拟系统。实验表明该方法是构建模糊入侵检测引擎的有效工具。

1 面向混合数据的自反馈模糊聚类分析

面向混合数据的自反馈模糊聚类分析是在基于摄动的模糊聚类算法(FCMBP算法)[4]的基础上改造而成,它有效地缓解了传统模糊聚类算法[5]存在的转换失真问题,在理论基础完备性和聚类结果准确性上表现更好。

1.1 基于摄动的模糊聚类方法

设定数据集 $X=\{x_1,x_2,\cdots,x_n\}\in R_n$ 是 n 维模式空间的一个特征数据集,根据相似性度量,该集合

收稿日期:2007-06-30
作者简介:令狐大智(1979-),男,高级工程师,主要从事信息安全、人工智能研究。
*广西科技攻关项目(桂科攻关0385001)、广西留学回国人员科学基金项目(桂科回0342001)资助。

可聚集成 m 个子集 $X_1,X_2,\cdots,X_m,2\leqslant m\leqslant n$，它们组成的特征向量为 X 的一个模糊划分，R_{ik} 表示特征数据 x_i 与特征数据 x_j 的模糊相似度，从而可得模糊相似矩阵

$$R=[r_{ik}]_{n\times n},$$

其中，$0\leqslant r_{ik}\leqslant 1(1\leqslant i\leqslant n,1\leqslant k\leqslant n)$。

令 φ_n 为 n 阶模糊相似矩阵全体，$R\in\varphi_n$，称 $Y=(y_{ij})n\times n$ 为 R 的摄动矩阵，若 Y 满足

$$y_{ij}\in[-r_{ij},1-r_{ij}],y_{ii}=0,y_{ij}=y_{ji},$$

记 $R+Y\cong(r_{ij}+y_{ij})_{n\times n}$，若 $(R+Y)^2=(R+Y)$，则 $R+Y$ 为模糊等价矩阵。进而 FCMBP 的目标函数可表示为

$$F(x)\triangleq\frac{1}{2}\|R-X\|_F^2=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_{ij}-r_{ij})^2,$$

其中 $\|\cdot\|_F$ 表示 Frobenius 范数。

最优模糊等价矩阵转为使 $F(x)$ 达到最小值的解 $x\in x_n$。

1.2 面向混合数据的自反馈模糊聚类分析算法

面向混合数据的自反馈模糊聚类分析算法的构造过程分四步完成，即：数据的前期处理，分析、构建相似矩阵 R，求解最优模糊等价矩阵 R^n，进行数据聚集。具体步骤如下。

步骤 1：分析数据记录中每个属性的特征及其入侵作用度，获取关键属性列表、属性作用度列表和属性类型列表，建立对应关系；

步骤 2：根据上述对应关系列表和相似性计算基础公式，构建基于作用度和属性关系的相似性计算公式，并构建相似矩阵 R；

步骤 3：根据相似矩阵 R 建立模糊等价矩阵 χ_n，对 χ_n 构造其解的参数表示得到参数图，进而获得 χ_n 的全体解集合，并求 χ_n 的基础解系，其中

$$\chi_n=\{(x_i(t_1,t_2,\cdots,t_{n-1}))_\sigma|1\leqslant i\leqslant k(n),\sigma\in\sum,t_j\in[0,1],1\leqslant j\leqslant n-1\};$$

步骤 4：对照 R 与 χ_n 的基础解系，从基础解系中寻找最相似者 x_i；

步骤 5：对 x_i 进行多次置换，使 x_i 与 R 更相近，并设计 σ_i；

步骤 6：计算 $F((x_i)_\sigma)$，求其最小值点，取 $R=(x_i(t_1^i,t_2^i,\cdots,t_{n-1}^i))_\sigma$，其中用 $F((x_i)_\sigma)$ 的结果来代替 $t_1,t_2,\cdots,t_n$，将其作为最优等价阵 R^n；

步骤 7：根据性能要求进行数据聚集，形成各个聚类集合。

2 算法测试实验

为了验证文中所提的算法性能，采用 KDD99[6] 数据集，结合基于传递闭包法的模糊聚类方法进行对比测试。

2.1 数据预处理

KDD99 数据集中数据属性包括字符型和数值两大类，区间变量、二态变量、分类变量和序数变量四小类，每类数据都有相应的度量标准和取值范围，在模糊聚类分析中，它们将对数据处理造成很大影响，比如局部优化问题。因此，针对属性类型将其进行标准化和归一化，消除类型差别所造成的影响。

(1)字符型变量离散化、序列化处理，使其转换为序列化数值型数据；

(2)采用公式 $Z_{ik}=(x_{ik}-f_k)/\sigma_k$ 对所有数据进行标准化，使所有取值限定在[-1,1]范围内，消除局部化影响，其中，f_k 为第 k 种属性取值的平均值；x_{ik} 为第 i 条记录的第 k 个属性取值；σ_k 为第 k 种属性取值的均方差。

(3)公式采用 $x_{ij}=(x_{ij}-x_{\min})/(x_{\max}-x_{\min})$ 进行归一化操作，使最终数据限定在[0,1]范围内，满足聚类分析的要求，其中 $x_{\min}$ 记录第 j 个属性取值中的最小值；$x_{\max}$ 记录第 j 个属性取值中的最大值。

2.2 模糊等价矩阵获取

根据数据分析阶段获得的关键属性列表、属性作用度列表和属性类型列表，在预处理的基础上计算两记录间的相似程度 r_{ij}。计算公式为

$$r_{ij}=Sim(e_i,e_j)=\alpha Sim^{(n)}(e_i,e_j)+\beta Sim^{(s)}(e_i,e_j),$$

其中 α,β 表示记录种数值型属性和字符型属性各自代表的权重；$Sim^{(n)}(e_i,e_j)$ 表示给定两个记录中数值型属性间的相似性，$Sim^{(s)}(e_i,e_j)$ 表示字符属性间的相似性。

其中 $Sim^{(n)}$ 和 $Sim^{(s)}$ 的计算公式为

$$Sim^{(n)}(e_i-e_j)=\sum_{k=1}^{m}(x_{ik}-\bar{x}_i(x_{jk}-\bar{x}_j)/\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2}\cdot\sqrt{\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2},$$

其中 $\bar{x}_i,\bar{x}_j$ 为平均值，m 为记录个数；

$$Sim^{(s)}(e_i-e_j)=\frac{m-a}{m},$$

其中 m 为每个对象的属性个数；a 为对象 ij 中取值相同的属性个数。

2.3 实验及其结果

根据本算法和传递闭包算法的差异，我们构造三组实验数据：test1，用于验证转换失真问题在入侵检测领域也存在，且 FCMBP 算法能够缓解该问题；

test2,按 KDD99 数据类型比例,随机抽取 2 万行数据集合,用于实际测试;test3,采用 test2 构造原理,抽取与 test2 不同的 10 万行数据,用于系统学习。

2.3.1 问题验证

利用本算法和传递闭包算法分别对 test1 数据集合进行模糊聚类的结果见表 1。

表 1 两种算法模糊聚类对比结果

本文提出的算法		传递闭包算法聚类算法	
阈值取值	聚类集合	阈值取值	聚类集合
1	{1}{2}{3}{4}{5}{6}{7}{8}{9}	1	{1}{2}{3}{4}{5}{6}{7}{8}{9}
0.82	{1}{2}{3}{4,9}{5}{6}{7}{8}	0.82	{1}{2}{3}{4,9}{5}{6}{7}{8}
0.80	{1,2}{3}{4,9}{5}{6}{7}{8}	0.80	{1,2}{3}{4,8,9}{5}{6}{7}
0.78	{1,2}{3}{4,8,9}{5}{6}{7}	0.79	{1,2,4,8,9}{3}{5}{6}{7}
0.663	{1,2}{3}{4,7,8,9}{5}{6}	0.74	{1,2,4,7,8,9}{3}{5}{6}
0.636	{1,2,4,7,8,9}{3}{5}{6}	0.69	{1,2,3,4,7,8,9}{5}{6}
0.523	{1,2,3,4,7,8,9}{5}{6}	0.14	{1,2,3,4,6,7,8,9}{5}
0.104	{1,2,3,4,56,7,8,9}{5}	0.13	{1,2,3,4,5,6,7,8,9}
0.088	{1,2,3,4,5,6,7,8,9}		

从表 1 可以得出:将模糊聚类引入入侵检测同样存在转换失真问题,并且影响到实际的检测结果,本算法的结果优于传递闭包算法的结果;在取阈值分别为 0.663 和 0.79,将记录仅看作正常/异常数据,则本算法的误报率为 0,而传递闭包算法的为 1/9,同样本算法的可解释性更强。

2.3.2 仿真验证

利用 test3 数据集合获取本算法算法和传递闭包算法下的适当阈值,用于 test2 数据集合进行实际测试的结果见表 2。

表 2 检测率结果对比

数据类型	本算法检测率(%)	传递闭包算法检测率(%)	本算法误报率(%)	传递闭包算法误报率(%)
Back	82.32	80.98	0.98	0.54
Pod	60.0	57	1.62	1.29
Guest	100	96	0.22	1.74
Portsweep	71.3	34.13	0.83	2.43
正常	98.7	99.1	1.35	1.33
未知	69	87		

从表 2 可以看出本算法在对 4 种入侵行为的检测率上都优于传递闭包算法,在误报率上针对 Back 和 Pod 类入侵相差不大,但在 Guest 和 Portsweep 上明显优于传递闭包算法,其平均检测率分别为 82.464%和 73.442,平均误报率分别为 1%和 1.466%,显然本算法优于传递闭包算法指导下的模糊聚类入侵检测引擎。但在进行检测的过程中仍发现需要进一步改进的地方,例如,入侵类型数据间检测误归类比率比较高,针对大规模数据集合检测处理速度过慢等。

3 结束语

本文所提出的面向混合数据的自反馈模糊聚类的入侵检测数据分析方法是在模糊聚类的基础上增加自反馈功能和数据属性对应分析功能,对网络数据集进行检测。通过对 KDD99 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。由于本算法的基础 FCMBP 聚类对初始化数据较敏感,并且最优等价阵的计算也期待更优良的算法,因此进一步的工作是要将神经网络、遗传算法等思想与本文所提聚类思想相结合,进一步降低入侵检测系统的误报率和漏报率,提高系统的自适应能力,改善系统的综合性能。

参考文献:

[1] AXELSSON S. Research in intrusion detection system: a survey [R], Chalmers University of Technology, CMU/SEI Technical Report (CMU/SEI-99-TR-028), 1999.

[2] 肖政宏,周学清,尹浩. 基于支持向量机和 Hurst 指数的集成入侵检测研究[J]. 微计算机信息,2006,22(9):51-52,64.

[3] HAI JIN, JIANHUA SUN, HAO CHEN, et al. A fuzzy data mining based intrusion detection model: proc of 10th Int Workshop on Future Trends of Distributed Computing Systems[R]. Suzhou, China, 2004: 191-197.

[4] LI HHONG-XING. Fuzzy clustering method based on perturbation[J]. FuzzySets and Systems, 1989, 33(3): 291-302.

[5] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.

[6] KDD CUP. KDD[EB/OL]. [2007-09-10]. http://kdd.ics.uci.edu/databases/kddcup98/ kddcup98.html.

(责任编辑:邓大玉)