第12卷 第23期 2012年8月
1671—1815(2012)23-5797-04

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 12 No. 23 Aug. 2012

基于聚类集成的网络入侵检测算法

赵 晖

(陕西理工学院数学与计算机科学学院,汉中 723000)

摘 要 为了进一步提高网络入侵检测的效果,提出一种基于聚类集成的入侵检测算法。首先利用Bagging算法从训练集中生成多个训练子集。然后调用模糊C均值聚类算法训练并生产多个基本聚类器。然后利用信息论构造适应度函数。采用粒子群算法从上述聚类集体中获得一个具有最优性能的集成聚类器。仿真实验结果表明,该算法能有效的提高入侵检测的精度,具有较高的泛化性和和稳定性。

关键词 入侵检测 聚类集成 Bagging 模糊C均值 粒子群算法

中图法分类号 TP393.08; **文献标志码** A

入侵检测是一种重要的主动网络安全技术,可有效发现来自网络外部与内部误操作引起的攻击。它与防火墙等静态安全技术配合使用,能有效提高网络的安全性。聚类是一种无监督的学习,是按照一定的要求和规律对物理或抽象对象的集合进行区分和分类的过程。它要求同类的对象具有某种共同内涵。聚类的这种特性,可以用来构建一种无监督、高效的入侵检测系统,成为人们研究的热点。建立此类检测模型的聚类分析方法主要有基于阈值的最近邻聚类算法、K-均值及其改进算法、随机聚类(EM)算法等[1]。但这些聚类方法是一种硬划分,将需要划分的对象严格的划分到某一种类中,是一种类别分明的划分方式。而入侵检测中,用户行为偏离正常行为模式的程度是个模糊概念。基于上述原因,文献[2,3]提出了模糊聚类进行入侵检测,有效的提高了检测效果。然而聚类算法很大程度受到相关参数及初始化的影响,特别不同的聚类算法得到的聚类结果往往具有较大的差异,这就使得选择合适的聚类算法是一件非常重要,同时也是一件很困难的工作。

为了解决这些问题,Strehl 等提出了聚类集成[4],利用集成学习技术[5—7],通过学习产生数据集的多个聚类结果,基于某种策略进行合成得到新的聚类结果。研究表明,聚类集成对任意形状和规模的数据聚类时,其性能优于单一的聚类算法,可以提高聚类算法的鲁棒性和稳定性,特别使用户避免选用不恰当的聚类算法的风险。通常聚类集成是利用投票进行合成,但因为存在类标签的统一问题,所以使得投票法使用起来非常困难。

基于上述分析,本文首先采用bagging技术生成多个样本子集并调用模糊C均值聚类算法训练生成多个基本聚类器,然后利用信息论构造适应度函数,采用粒子群算法从聚类集成中获得一个具有最佳性能的集成聚类器。并通过仿真实验验证该集成算法在入侵检测中的有效性和稳定性。

1 聚类集成描述

聚类集成就是首先对一组对象产生多个聚类结果,然后采用某种策略将多个不同结果进行合成。聚类集成可以描述:假设数据集 $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$。

(1)对数据集 X 使用 N 次聚类算法,则得到 N 个聚类 $P = \{P_1, P_2, \cdots, P_N\}$,其中 $P_i(i = 1,2,3,\cdots,N)$ 为第 i 次聚类算法得到的聚类结果。

(2)采用某种策略(如投票法等)对 P 中的聚类结果进行合成得到一个新的数据划分 $p¢$。

(3)利用 $p¢$ 对新的样本数据进行分类判别。

2012年5月9日收到 陕西省教育厅科研基金2010jk459、12JK0864)和陕西理工学院科研基金(SLGKY11-08)资助

作者简介:赵 晖(1979—)男,硕士,讲师. 研究方向:数据挖掘、网络入侵检测。E-mail:zh911@sina.com.cn。

2 基于聚类集成的入侵检测算法

2.1 基本聚类器的产生

Bagging[8]是基于重采样技术的一种集成算法，各训练子集由从原始训练集中随机选取若干样本组成，训练集的规模通与原始训练集相当，训练样本允许重复选取。这样原训练集中某些样本可能在新的训练子集中出现多次，而另外一些样本可能一次也不出现。

为了生成具有较大差异度的聚类器，首先应增加训练集的差异性，本文采用 Bagging 算法生成多个训练子集，在每个训练子集上调用 FCM 算法训练并生成基本聚类器。这样得到的基本聚类器就具有较大的差异性，可以提高系统的泛化性能。

2.2 基本聚类器的合成

为了从所产生的聚类集体中找到一个与此聚类集体最“统一”的一个聚类结果，即聚类集成结果。本文利用信息论构建一个度量“统一”性的准则函数[9]。用与所有聚类成员的距离和作为准则函数，并以此函数为适应度函数调用 PSO 算法找到一个聚类使之与聚类集体中的所有基本聚类的距离和最小。

k 维单形体：

$SIMPLEX_{k-1} = \{(p_1,\cdots,p_k) \in R^k \times | \ p_i \geqslant 0$ and $p_1 + \cdots + p_k = 1\}$。

函数 $f:R\to R$ 在集合 $S \subseteq R$ 上是凹的，如果对任意的 $a \in [0,1]$ 和 $x,y \in R$ 满足：

$f(ax + (1-a)y) \geqslant af(x) + (1-a)f(y)$。

函数 f 在集合 R 上是可加的，如果它满足：

$f(x+y) \leqslant f(x) + f(y)$ $x,y \in R$。

定义 1[9] 发生器是一个凹函数和次可加函数 $f:[0,1]\to R$，对于任意 $(p_1,\cdots,p_n) \in SIMPLEX_{n-1}$ 和 $\theta \in [0,1]$ 它满足 $f(0) = f(1) = 0$ 和 $f(\theta p_1) + \cdots + f(\theta p_n) \leqslant \theta(f(p_1) + \cdots + f(p_n) + f(\theta))$。

本文使用发生器：

$f_{ge}(p) = -(p - p^2)\lg(p - p^2)$。

定义 2[9] 假设 f 是一个，$\pi = \{B_1,\cdots,B_n\}$，$\sigma = \{C_1,\cdots,C_m\}$ 是数据集 S 的两个聚类。则聚类 π 的 f 熵（$f-$entropy）定义：$H^f(\pi) = f(\frac{|B_1|}{|S|}) + \cdots + f(\frac{|B_n|}{|S|})$；

子集 $L \subseteq S$ 相对于聚类 π 的 f 熵定义：$IMP^f_\pi(L) = |L|\left(f\left(\frac{L\cap B_1}{|L|}\right) + ,\cdots, + f\left(\frac{L\cap B_n}{|L|}\right)\right)$；

子集 $L \subseteq S$ 相对于聚类 π 的特定 $f-$impurity 定义：$imp^f_\pi(L) = \left(f\left(\frac{L\cap B_1}{|L|}\right) + ,\cdots, + f\left(\frac{L\cap B_n}{|L|}\right)\right)$；

π 相对于 σ 的条件 $f-$entropy 定义：

$$H^f(\pi|\sigma) = \sum_{j=1}^{m}\frac{|C_1|}{S} imp^f_\pi(C_j) = \frac{1}{|S|}\sum_{j=1}^{m}|C_j| \times \sum_{i=1}^{n} f(\frac{B_1\cap C_1}{|C_j|});$$

$H^f(\pi|\sigma)$ 是聚类 σ 的类相对于聚类 π 的特定 f-impurity 的平均值。

定义在集合 S 上的距离是一个函数 $d:S\times S\to R$，对于任意的 $x,y \in S$，满足 $d(x,y) \geqslant 0$，$d(x,x) = 0$ 和 $d(x,y) = d(y,x)$。使用广义熵和广义条件熵可以定义两个聚类的距离函数。假设 f 是一个发生器，π,σ 是数据集 S 上的两个聚类，则它们间的距离定义：$d^f(\pi,\sigma) = H^f(\pi|\sigma) + H^f(\sigma|\pi)$。当聚类 π 和 σ 相似时，意味着它们的类有许多相同的元素，所以 $H^f(\pi|\sigma)$ 和 $H^f(\sigma|\pi)$ 都接近于 0，它们间的距离 $d^f(\pi,\sigma)$ 也就接近于 0。

所以，在给定一个聚类集体 $\Pi = \{\pi_1,\pi_2,\cdots,\pi_N\}$ 上，最终聚类集成的结果 π^* 的准则函数定义为：

$$M_f(\pi^*) = \sum_{i=1}^{N} d^f(\pi^*,\pi) = \sum_{i=1}^{N} H^f(\pi^*|\pi_1) + H^f(\pi^*|\pi_1)。$$

本文利用 PSO 算法找到一个合适的聚类 π^*，使得 M 取得最小值。

输入：训练集 $D = \{(x_1,y_1),\cdots,(x_n,y_n)\}$，基本聚类器 $h_t(t = 1,2,3,\cdots,T)$，T 为最大迭代次数。

输出：集成聚类器 π^*

（1）种群的初始化：

①编码方式：每个聚类器是一个粒子，用每个数据点的簇标签形成的一个数字串来表示一个聚类或者说一个粒子，其中第 i 个数字表示相应聚类给予 x_t 的类标签。

②初始种群产生:随机选取 n 个位于区间(1, k]的整数,组成一个长度为 n 的数字串,即形成一个粒子,并给该粒子随即生成一个速度,多次重复该操作,则产生一个粒子种群。

(2)对每个粒子,利用准则函数计算适应度并和它经历过的最好位置的适应度进行比较,如果适应度值更小,则更新该粒子的最好位置;

(3)对每个粒子,比较它的适应度和群体所经历的最好位置的适应度,如果适应度值更小,则更新全局最好位置;

(4)根据下面两式调整粒子的速度和位置:

$$V_{id} = WV_{id} + c_1 rand()(P_{id} - X_{id}) + c_2 Rand() \times (P_{gd} - X_{id}); X_{id} = X_{id} + V_{id}。$$

(5)计算新一代粒子的适应度值,若适应度值不再明显的减小或达到最大迭代次数,则算法终止;否则转入步骤(2)继续。

3　仿真实验

3.1　实验数据

我们选取 KDD CUP 1999 数据集中的 10% 数据集作为实验数据集。10% 数据集包括训练集和测试集两部分,从 10% 数据集的训练集中随机抽取 1 5000个样本作为训练数据 A,其中包含入侵数据 1 500个,其中包含四大类攻击 11 种,具体为 Dos 攻击有 ueptne 攻击 210 个、smurf 攻击 500 个、back 攻击 25 个,probing 攻击有 ipsweep 攻击 50 个、portsweep 攻击 50 个、satan 攻击 80 个,U2R 攻击有 buffer_overflow 攻击 15 个、rootkit 攻击 10 个,R2L 攻击有 waremster 攻击 10 个、Guess_passwd 攻击 20 个、warezclient 攻击 30 个。从 10% 数据集的训练集中随机抽取 10 000 个样本分别作为作为测试数据 B 其中包括攻击 1 000 个。

为验证本文算法对未知攻击的检测能力,在测试数据集 B 中加训练集 A 中没有的攻击类型,分为 4 大类 10 种,分别为 Dos 攻击有 teardrop20 个、pod40 个、land10 个,probing 攻击有 nmap20 个,U2R 攻击有 loadmodule3 个、perl2 个,R2L 攻击有 spy2 个、phf4 个、imap10 个、multihop4 个、ftp_write4 个,测试集数据共包含攻击类型 21 种。

3.2　算法评价标准

$$\text{检测率} = \frac{\text{正确检测出的攻击样本数量}}{\text{总的攻击样本数量}};$$

$$\text{误警率} = \frac{\text{被错误判断为攻击的正常样本数量}}{\text{总的正常样本数量}}。$$

3.3　实验过程

3.3.1　特征提取

KDD CUP 1999 数据集每条记录包含从一条连接中提取的 41 个属性值,属性之间具有信息重叠和冗余及噪声,为了获得较高的检测效果,本文采用文献[7]提出的 CFS[10] 方法进行特征提取,得到一个包含 10 个属性的特征子集,在剔除冗余属性的同时达到降维的目的,既提高检测精度,又提升时间效率。

3.3.2　算法参数设置

模糊 c – 均值聚类允许最小误差为 10^{-7},平滑因子 $m = 2$;粒子群群体为 50,设最大迭代次数为 400,平滑因子 $c_1 = c_2 = 1.5$,惯性权重 $W_{\max} = 0.8$, $W_{\min} = 0.6$。

3.3.3　实验过程与结果分析

将该算法在数据集 A 上进行训练,并分别在 B 上进行测试,实验重复 10 次,取其平均值作为实验结果。并与算法 1(基于 *FCM* 的入侵检测)和算法 2(*FCM* + *Bagging* + 投票法)的结果进行比较,具体结果见表 3.1—表 3.2。

表 3.1　不同算法在整个测试集上的检测结果

	算法 1	算法 2	本文算法
检测率	82.71%	86.24%	93.33%
误警率	3.53%	3.31%	3.07%
标准差	0.125 1	0.097 6	0.068 9

表 3.2　不同算法对已知攻击、未知攻击的检测结果

攻击类型	已知攻击			未知攻击		
	算法 1	算法 2	本文算法	算法 1	算法 2	本文算法
DoS	82.03%	90.14%	93.28%	76.73%	84.55%	92.29%
Probing	80.96%	89.32%	91.91%	78.13%	86.69%	91.01%
U2R	88.11%	93.56%	96.22%	83.78%	89.63%	95.16%
R2L	87.01%	92.89%	95.68%	81.32%	85.59%	94.21%

由表 3.1 数据可以看出,本文算法和算法 2 的检测率明显高于算法 1,说明利用聚类集成要比单一聚类结果更加准确。本文算法比算法 2 的精度平

均提高 6%,表明基于信息论的适应度函数调用 PSO 可以获得优于投票法的结果,因为 PSO 可以获得全局最优解,而且加快搜索速度。从表 2 数据看到,本文算法相对于算法 1 对已知攻击和未知攻击检测率都有大幅度的提高,对于已知攻击平均提高约 10%,而对于未知攻击平均提高约 15%,充分说明了该算法对未知攻击检测的有效性,算法具有较强的泛化性能和鲁棒性。

另外,标准差是衡量算法稳定性的重要指标,标准差越小算法越稳定。从表 1 数据中我们清楚的看到,本文算法的标准差最小,代表了算法的稳定性越高,因为聚类集成可以减小参数及初始化对聚类结果的影响。

4 结论

考虑到单一聚类结果的差异性以及泛化性能低等问题,本文提出基于聚类集成的入侵检测算法,在 Bagging 基础上利用 PSO 获得最优集成聚类器。仿真实验取得了较好的检测效果,特别是该算法对于含有未知攻击的检测具有更强的泛化性能,为入侵检测提供了一种有效的方法。

参 考 文 献

1 肖 敏,韩继军,肖德宝,等. 基于聚类的入侵检测研究综述. 计算机应用,2008;28(s1):34—42

2 杨德刚. 基于模糊 C 均值聚类的网络入侵检测算法. 计算机应用,2005;32(1):86—91

3 鲜继清,郎风华. 基于模糊聚类理论的入侵检测数据分析. 重庆大学学报,2005;28(7):74—75

4 Strehl A, Ghosh J. Cluster ensemble-a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 2003; 3(3):583—617

5 陈 涛. 基于双重扰动的支持向量机集成. 计算机应用,2011;28(1):46—49

6 陈 涛. 基于加速遗传算法的选择性支持向量机集成. 计算机应用研究,2011;32(2):57—61

7 陈 涛. 选择性支持向量机集成算法. 计算机工程与设计,2011;(05):259—263

8 Bagging B L. Predictors. Machine Leaming,1996;24(2):123—140

9 罗会兰. 聚类集成关键技术研究. 杭州:浙江大学计算机学院博士学位论文,2007;85—87

10 Mark A H. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of 17th International Conference on Machine Learning,2000;359—366

Network Intrusion Detection Algorithm Based on Clustering Ensemble

ZHAO Hui

(School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723000, P. R. China)

[**Abstract**] To improve the ability of network intrusion detection, a detection algorithm based on clustering ensemble is presented. First, many training subsets were produced from training dataset by Bagging, and clustering individuals were trained by fuzzy c-means clustering. Then, fitness function was construct using information theory, ensemble clustering machine of better ability were obtained from clustering individuals based on particle swarm optimization algorithm. The experiments show that the algorithm effectively improve accuracy of intrusion detection, it have higher generalization performance and stability.

[**Key words**] network intrusion detection clustering ensemble Bagging fuzzy c-means clusting (FCM) particle swarm optimization algorithm(PSO)