基于模糊C均值聚类的网络入侵检测算法*)

杨德刚

(重庆大学计算机科学与工程学院　重庆400030)　(重庆师范大学数学与计算机科学学院　重庆400047)

摘　要　入侵检测已成为网络安全的第二层重要防御线。分析了对新型未知的攻击的入侵检测，提出基于模糊C均值聚类的网络入侵检测算法。用KDD-99数据集的仿真实验结果表明算法的可行性、有效性和可扩展性，并有效提高了聚类检测的检测率，降低了误检率。

关键词　入侵检测，聚类分析，模糊C均值聚类

Research of the Network Intrusion Detection Based on Fuzzy Clustering

YANG De-Gang

(College of Computer Science and Engineering, Chongqing University, Chongqing 400030)

(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047)

Abstract　The intrusion detection become the second floor defense line of the network security. Analyze to the characteristic of the intrusion detection technique for newly and unknown attack, and brings forward algorithm of Network Intrusion Detection based on Fuzzy C-means Clustering. The result of simulations run on the KDD-99 datasets show the feasible, efficient and extensible for unknown intrusion detection, and increase detection rate of the clustering detection and decrease the false alarms rate.

Keywords　Intrusion detection, Clustering analysis, Fuzzy c-means clustering algorithm

1　引言

入侵检测已成为传统安全技术，如防火墙之后的第二道安全防御线。入侵检测从技术上主要分为误用检测和异常检测两类。前者通过特征匹配检测是否发生入侵，对已知入侵检测准确，速度快，但它不能检测未知入侵；而后者则在建立正常模型基础上，通过实际行为与之比较是否偏离来判断入侵行为的发生。它能检测已知入侵，也能检测未知入侵。因此异常检测成为当前研究的热点领域。

聚类分析用于发现数据实例中的隐性模式和用于检测入侵中有意义的特征。好的聚类方法是要将一个数据集划分为若干组，并且具备高的组内相似性和低的组间相似性两个特点。因此，聚类分析用于入侵检测是可行的。Portnoy[3]最先提出基于聚类分析的入侵检测技术，该技术在对数据实例进行正规化处理，然后采用欧氏距离，并用单链法进行聚类，经过标识，通过分类以检测入侵。其检测率可达50%，误检率1%左右。该方法存在自适应和可扩展性不强的特点。本文则根据聚类和入侵实例的特征，提出基于模糊C均值聚类的入侵检测算法，仿真实验结果表明该算法能有效提高检测率，降低误检率。

2　基于模糊聚类的入侵检测算法

2.1　基于聚类分析的入侵检测过程

用聚类分析进行入侵检测其过程如下：

(1)*收集实例数据*　网络入侵检测系统首先收集网络上传输的数据流，通常用抓包工具来收集网络数据包，其中最常用的如UNIX下的TCPDUMP，它能够监听和接收网络中所有正在传输的数据包，并把它们记录到文件中。原始的网络数据包并不适合进行聚类分析，需要将原始的网络数据包恢复成TCP/IP层的连接记录，其中每个连接记录代表一次TCP/IP层的连接，并包含一个网络连接的多个属性，如网络协议、连接起始时间、连接结束时间、服务、源IP地址、目的地址、连接终止状态等。

(2)*实例数据标准化*　由于采集的网络数据的属性值之间的差别可能很大，而且它们可能用不同的单位来度量。为了消除由于度量不同对聚类的影响，应对测量的属性值进行标准化，方法如下：

$$x'_{ij}=\frac{x_{ij}-\overline{x_j}}{S_j},\begin{pmatrix}i=1,2,\cdots,n\\ j=1,2,\cdots,m\end{pmatrix},\qquad(1)$$

其中$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$S_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}$

x'_{ij}即为标准化后的实例数据值。通过计算每个特征值与平均值之间的标准偏差，可得到该特征值在正规化空间中的新值。

(3)*数据初始化聚类*　对收集的连接记录进行标准化，即可进行初始化的聚类过程。能够利用的聚类算法有很多，如可以采用层次的聚类算法和基于模型的聚类方法等。

(4)*对初始化聚类集的优化*　为了获得更优化的聚类结果，应对初始化的聚类簇进行优化，以尽可能地将相似的实例聚为一类，相异的实例分到不同聚类。

(5)*聚类簇的划分*　由于正常条件下正常实例占网络行为的大部分，因此可以假设正常行为在数量上远远大于各种攻击行为。当使用聚类算法得到聚类簇后，即利用这一假设进行划分，区分哪些是正常连接数据，哪些是入侵连接数据。

*)基金项目：重庆师范大学校内基金(计算机网络安全)。杨德刚　硕士生，讲师，主要研究方向：数据挖掘、信息安全。

当聚类簇完成后，可将所有的聚类按其中包含的实例数量大小进行排序，设定阈值 m，对大于 m 的聚类簇即认为是正常行为，否则即为异常。

2.2 初始化聚类检测

在聚类入侵检测过程中，进行初始化聚类的算法步骤如下：

1)初始化空聚类集合；

2)取标准化后的数据集中第一条连接数据建立第一个聚类并保存到类集合；

3)每取出一个新的连接实例 I_i，都计算其与各个已存在的聚类集合的聚类中心距离。取其中最小的距离 $d_{\min}=\mathrm{MIN}(d_{I_i},C_i)$，其中 $i\in 1,\cdots,k$，d_{I_i,C_i} 为实例 I_i 与聚类的距离。k 为当前类集合中已建立的聚类个数；

4)判断 $d_{\min}$，若它 $d_{\min}\leqslant d_0$，d_0 为事先定义的聚类宽度值，则将其放入已存在的一种聚类簇，然后继续处理下一条实例数据；否则如果当前实例与聚类簇集合中所有聚类实例的距离都大于 d_0，则以该实例为中心定义一个新的聚类，并加入到聚类簇集合中；

5)当训练样本集中的所有数据都处理完，则结束，否则转3)。

通过以上的聚类算法后，将得到一个经过初步聚类后的聚类簇。

2.3 基于模糊 C 均值聚类的入侵检测算法

2.3.1 *模糊 C 均值聚类*(FCM) 是一种基于划分的聚类算法，是普通 C 均值算法的改进，普通 C 均值算法对于数据的划分是硬性的，而 FCM 则是一种柔性的模糊划分。模糊 C 均值聚类算法适用于具有椭球状分布的数据聚类。FCM 以所分组内平方和最小化为判据，用隶属度确定每个数据实例属于某个聚类的程度，把 n 个数据实例 $X=\{X_i|X_i\in R(i=1,2,\cdots,n)$ 分为 K 类($1<K<N$)，并求每组的聚类中心，使得非相似性指标的价值函数达到最小。其分类矩阵 $U=(u_{ij}|i=1,2,\cdots,n;j=1,2,\cdots,k)$，其中元素 u_{ij} 表示第 i 个数据实例属于第 j 类的隶属度，并满足如下条件：

$$\sum_{j=1}^{k}u_{ij}=1,\forall\ i=1,\cdots,n \tag{2}$$

FCM 进行模糊划分，使每个给定数据实例在区间0,1之间的隶属度确定其属于各个组的程度。隶属矩阵 U 的元素在0,1之间取值。定义价值函数为：

$$J_m(U,C)=\sum_{i=1}^{N}\sum_{j=1}^{k}u_{ij}^m d_{ij}^2(X_i,C_j) \tag{3}$$

J_m 可看作是每个数据实例到类中心的距离平方和。式中，$C_i\in I$ 为类中心，$C=\{C_j|C_j\in I,j=1,2,\cdots,k\}$；$X_i\in I$ 为数据实例数据集；u_{ij} 是第 I 个数据实例属于第 j 个聚类中心的隶属度，取值在[0,1]间，$U=\{u_{ij}\}$ 是一个 $n\times K$ 矩阵，$C=[C_1,C_2,\cdots,C_k]$ 是一个 $s\times k$；$C=\{c_1,c_2,\cdots,c_k\}$，c_i 为模糊组 i 的聚类中心；$X_i\in R^p$ 为数据实例；$d_{ij}(X_i,C_j)$ 为第 i 个数据实例与第 j 个聚类中心之间的欧氏距离；m 是模糊系数($1\leqslant m<\infty$)；k 为预分类组数，由初始聚类确定。应用拉格朗日乘数法，使式(3)得到最小值的必要条件为：

$$u_{ij}=1/\sum_{l=1}^{k}(d_{ij}/d_{il})^{2/(m-1)},\forall\ i \tag{4}$$

$$c_{ij}=(\sum_{i=1}^{n}u_{ij}^m x_j)/(\sum_{i=1}^{n}u_{ij}^m),\forall\ j \tag{5}$$

式中参数 m，是控制算法的标量，用来控制分类矩阵 U 的模糊程度，m 越大越模糊。如果 $m=1$ 则 FCM 算法退化为 HCM 聚类算法。FCM 聚类需要多次进行迭代计算，以使价值函数取得最小值。

2.3.2 *基于 FCM 的入侵检测算法* 从上面的讨论可知，FCM 算法需要两个参数：聚类数目 C 和参数 m。聚类数目利用初始聚类结果中的聚类数作为 C，这里的 C 远远小于聚类样本的总个数。用初始聚类算法来确定初始聚类中心和聚类数量。基于 FCM 的入侵检测优化阶段按以下步骤进行：

步骤1 用值在0,1间的随机数初始化隶属矩阵 U，并使其满足式 $\sum_{i=1}^{n}u_{ij}=1,\forall\ j=1,\cdots,n$ 中的约束条件。

步骤2 用式 $c_i=(\sum_{j=1}^{n}u_{ij}^m x_j)/\sum_{j=1}^{n}u_{ij}^m$ 计算 k 个聚类中心 c_i，$i=1,\cdots,k$。

步骤3 用 $u_{ij}=1/\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}$ 计算新的隶属度矩阵 U。

步骤4 根据式 $J_m(U,C)=\sum_{i=1}^{N}\sum_{j=1}^{k}u_{ij}^m d_{ij}^2(X_i,C_j)$ 计算价值函数。如果它小于某个确定的阈值，或它相对上次价值函数值的改变量小于某个阈值，则算法停止，并输出聚类结果。否则，返回步骤2继续迭代。

算法的输出是 $N\times K$ 的一个模糊划分矩阵，这个矩阵表示的是每个样本点属于每个类的隶属度。根据这个矩阵按照模糊集合中的最大隶属原则就能够确定每个样本点归为哪个类。从算法过程可知，算法对于满足正态分布的数据聚类效果较好，对孤立点也比较敏感。

3 实验结果分析

为评价提出基于 FCM 聚类的入侵检测算法，选用 KDD Cup 1999网络数据集[7]。该数据集由麻省理工学院 Lincoln 实验室仿真美国空军局域网环境而建立的网络流量测试数据集。数据集包含了7个星期网络流量，大约500万条连接记录。这些记录中包括多种广泛的网络环境下的模拟入侵，其中每个连接实例包含42个属性而且均已标识为正常或特定的攻击类型。这些数据集包括22种攻击和1个正常类型。连接数据集中的入侵类型按攻击手段可大致分为：DOS、R2L、U2R、Probing 四类。数据集中的数据格式如下：

0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,254,1.00,0.01,0.00,0.00,0.00,0.00,0.00,0.00,snmpgetattack

在这条数据中，第一个属性是连接时间；第二个属性表明该连接是 TCP 还是 UDP 数据包；第三个属性表示服务类型，如 http、ftp、smtp 等；第四个属性表明连接标记，如 SF、REJ、RSTR 等；随后的37个为该连接的数值属性，即连接时的记录参数；最后一个属性是类标记属性，表明这条记录是正常连接还是入侵连接。

为方便仿真实验，采用随机抽取的5组2100个样本数据集。这是从 KDD Cup'99数据集中随机选取的。由于该数据集的仿真数据比较分散，子集中攻击数量和类型分布不平衡，导致某些子集仅包含某一种攻击类型，这对于反映整个网络真实环境是不利的，所以我们以随机方式重新建立5个数据集，使得每个数据集中有2000个正常连接，100个异常连接，而且使异常连接的种类尽可能平均。在实现聚类算法过程中忽略

(下转第91页)

Conf. 1995
9 Park J, Sandhu R. Towards Usage Control Models: Beyond Traditional Access Control. SACMAT02, Monterey, Califomia, USA, ACM, 2002
10 Park J, Sandhu R. Originator Control in Usage Contrlol. In: 3rd International Workshop on Policies for Distributed Systems and Networks(Policy02). June 2002
11 Sandhu R, Park J. Usage Control: A Vision for Next Generation Access Control. MMM-ACNS 2003
12 Sandhu R, Coyne E J, Feinstein H L, Youman C E. Role-based access control models. IEEE Computer, 1996, 29(2): 38～47
13 Sandhu R. Rationale for the RBAC96 family of access control models. In: Proc. of the 1st ACM Workshop on Role-Based Access Control. ACM, 1997
14 Barkley J. Comparing Simple Role Based Access Control Models and Access Control Lists. In: Proc. of the Second ACM Workshop on Role-Based Access Control, Nov. 1997. 127～132
15 Park J, Sandhu R, Schifalacqua J. Security Architectures for Controlled Digital Information Dissemination. ACSAC, 2000
16 Sandhu R. Access Control: The Neglected Frontier. ACISP, 1996
17 Chen F, Sandhu R. Constraints for RBAC, ACM RBAC, 1995
18 Thomas R, Sandhu R. Conceptual Foundations for a Model of Task-based Authorizations, CSFW, 1994
19 Sandhu R. Design and Implementation of Multilevel Databases. RADC, 1994
20 Sandhu R, et al. Role-Based Access Control: A Multi-Dimensional View. ACSAC, 1994
21 Thomas R, Sandhu R. Discretionary Access Control in Object-Oriented Databases: Issues and Research Directions. NCSC, 1993
22 Jajodia S, Sandhu R. Toward a Multilevel Secure Relational Data Model. SIGMOD, 1991
23 Sandhu R. Evaluation by Parts of Trusted Database Management Systems. RADC, 1991
24 Sandhu R, Share M. Some Owner-Based Schemes with Dynamic Groups in the Schematic Protection Model. OAKLAND, 1986
25 Sandhu R, Bhamidipati V, Munawer Q. The ARBAC97 Model for Role-Based Administration of Roles. ACM Transactions on Information and System Security, 1999, 2(1): 105～135
26 Park J, Sandhu R. The UCONABC Usage Control Model. ACM Transactions on Information and System Security (TISSEC), Feb. 2004
27 Yovits M C. Database Security. Advances in Computers. Academic Press, 1994, 38: 1～74

(上接第87页)

类标识属性，其仅供算法结果分析。表1和图1是各数据集中攻击类型分布情况。

表1 实验数据集结构

	实例数	正常实例数	入侵实例数	攻击类型数据	分类攻击数
数据集1	2100	2000	100	21	4
数据集2	2100	2000	100	11	3
数据集3	2100	2000	100	13	4
数据集4	2100	2000	100	8	2
数据集5	2100	2000	100	13	4

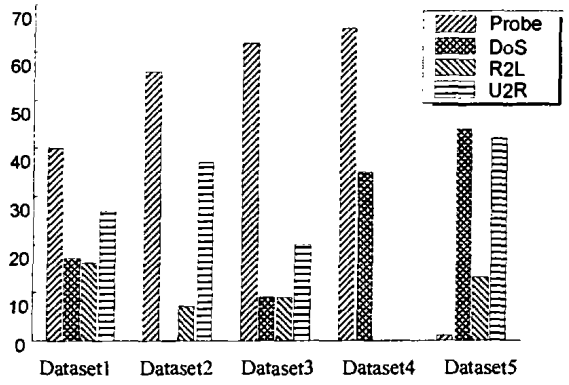

图1 实验数据集中攻击类型分布

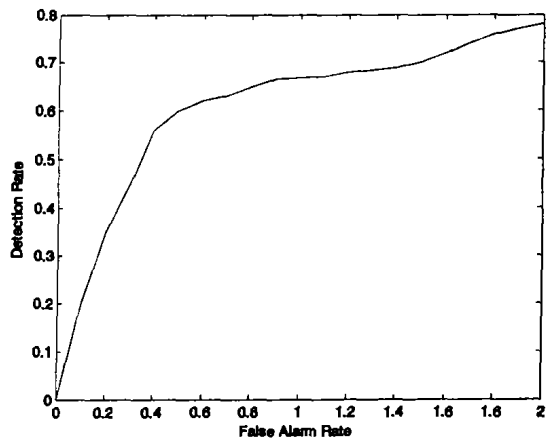

图2 NIDBFCM 检测率和误检测率的相对曲线

针对以上数据集，选用实验平台 Windows2000，Matlab6.5语言编程环境，Intel Celeron 1.70GHz CPU，256MB 内存。为评价入侵检测算法性能使用两个指标：检测率 DR(Detection Rate)和误检率 FR(False Alarm Rate)。检测率 DR 是被系统检测到的入侵实例数目与检测数据集中总的入侵实例数目之比；误检率 FR 则为被误判为入侵的正常实例数目与总的正常实例数目之比。这两项性能指标能充分反映算法的入侵检测能力。最好的入侵检测系统应该使 DR 尽可能的大，而 FR 则尽可能的小。图2是上述5组实例数据的平均检测结果，在隶属度改变的情况下，检测率提高的情况下，误检率也会提高。

从图2可知最好性能当 FR＝1%时，DR＝69%，这充分表明算法对于未知入侵行为检测的可行性和有效性，与 Portnoy 提出的方法相比，在一定程度上提高了检测率，降低了误检率。

结束语 本文分析了聚类技术及其在入侵检测中的应用，提出以 FCM 进行入侵检测，分析了检测过程，并用 KDD-99实验数据集进行仿真实验，其结果表明该算法提高了检测的效率，算法是可行的、有效的。

参 考 文 献

1 Wenke L. A Data Mining Framework for Building Intrusion Detection Models. In IEEE Symposium on Security and Privacy, 1999
2 Lee W, Stolfo S J, Mok K W. Mining in a data-flow environment: experience in intrusion detection. submitted for publication, 1999
3 Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering. In: Proc. of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001). Philadelphia: ACM Press, 2001(11)
4 Li Xiangyang. Clustering and Classification Algorithm for Computer Intrusion Detection: [PhD.]. Arizona State University, 2001.12
5 Equihua M. Fuzzy clustering of ecological data J. Ecol, 1990, 78: 561～567
6 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html