文章编号：1001-2486（2002）02-0059-05

一种应用聚类技术检测网络入侵的新方法*

梁铁柱[1,2]，李建成[3]，王 晔[1,2]

（1. 解放军理工大学通信工程学院，江苏 南京 210016； 2. 总参 61 所，北京 100039；3. 国防科技大学训练部，湖南 长沙 410073）

摘 要：基于聚类技术提出了一种能处理不带标识且含异常数据样本的训练集数据的网络入侵检测方法。对网络连接数据作归一化处理后，通过比较数据样本间距离与类宽度 W 的关系进行数据类质心的自动搜索，并通过计算样本数据与各类质心的最小距离来对各样本数据进行类划分，同时根据各类中的样本数据动态调整类质心，使之更好地反映原始数据分布。完成样本数据的类划分后，根据正常类比例 N 来确定异常数据类别并用于网络连接数据的实时检测。结果表明，该方法有效地以较低的系统误警率从网络连接数据中检测出新的入侵行为，更降低了对训练数据集的要求。

关键词：聚类；入侵检测；检测率；误警率

中图分类号：O235；TP391 文献标识码：A

A Novel Clustering-Based Method to Network Intrusion Detection

LIANG Tie-zhu[1,2], LI Jian-cheng3, WANG Ye[1,2]

(1. School of Communication Engineering, PLA Univ. of Science and Technology, Nanjing 210016, China; 2. The 61st Research Institute, General Staff, Beijing 100039, China; 3. Training Department, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Researchers have developed two general categories of intrusion detection, i. e. misuse detection and anomaly detection, which differ at model construction. Signature-based misuse detection, which can detect the well-known attacks, will do nothing when new attack comes. Even traditional anomaly detection can catch some new attacks, the learning process overly relying on the training data sets which contain either purely clean normal data or correctly labeled data makes it useless in most cases. To solve such a problem, a novel clustering based method, capable of proces sing training data sets without type label and/or containing unknown intrusion data, is presented in this paper. After the normalization of network connection data, cluster centroids which is null at first can be obtained gradually and automatically through comparing the distance between data instances and the predefined cluster width , and each data instance can be then classified into the cluster which has the minimum distance with it. To ensure that the clusters can best represent the data distribution, cluster centroids also can be dynamically adjusted according to data instances contained in this cluster. With the classified data instances, the anomaly data clusters can be easily identified using normal cluster ratio , therefore performing the real – time detecting of each real network connection datum. Experiment result shows that this method can not only detect some new attacks, from network connection data sets, with low false positive rate, but also tolerate more general data sets.

Key words: clustering; intrusion detection; detection rate; false positive rate

在网络安全问题日益突出的今天，如何迅速、有效地发现各类新的入侵行为，对于保证系统和网络资源的安全显得十分重要[2]。

滥用检测根据已知的系统和应用软件的弱点及其攻击模式的特征进行编码，并通过与审计数据的匹配来检测入侵。滥用检测具有较低的误警率，但它不能检测出新出现的一些入侵行为，故漏报率也较高。异常检测通过对审计数据的训练学习，从中发现正常使用行为模式，以定量的统计方式描述可接受的行为特征，并由测试数据和正常行为模式的偏差捕获到异常，以区分非正常的、潜在的入侵行为。通过检查与正常行为相违背的行为，异常检测能够发现一些新的未知的入侵[3]。但由于其正常行

* 收稿日期：2001-11-27
基本项目：国家“九七三”重点基础研究发展规划资助（G19980305084）
作者简介：梁铁柱（1974—），男，工程师，博士生。

为模型的建立完全依赖于对训练数据集中正常数据样本的学习[4]，所以保证该数据集的洁净性，即不包含任何异常数据，对建立一个实用的入侵检测系统是至关重要的。而实际上，要为系统的学习收集这样一个洁净数据集往往是不太容易的，一旦有入侵数据被作为正常数据出现在训练数据集中，必然导致该类入侵行为及其变种都将被系统视为正常数据。

在实验中，我们将一种基于聚类的入侵检测方法应用于 KDD Cup 1999 Data [6] 中。该数据集提供了从一个模拟的局域网上采集来的9个星期的网络连接数据，其训练数据集包含5百万个连接数据，测试数据集包含了2百万个连接数据。在训练数据集上，该系统获得了0.672的检测率，而误警率则仅为0.008，在测试数据集上的性能也基本接近。与传统的方法相比，以上结果充分说明了该方法的可行性与实用性。

1 应用聚类技术检测入侵

聚类可以用来发现数据中的隐性模式和用于检测入侵的有意义的特征。聚类的目的是要将一个数据集划分为若干组，使得组内的相似性大于组间相似性[7]。实现这样一种划分需要一个相似性度量，即取两个输入向量，返回反映这两个向量间相似性的数值。由于大多数相似性度量对输入向量中元素的值域非常敏感，因此每个向量都必须归一化，即令其在单位区间［0，1］上取值。

1.1 归一化处理

对数据集应用聚类算法前，必须将所考虑的数据集归一化到单位超立体中。对于给定的训练数据集，由(1)和(2)式分别求特征向量的均值 mean _ vector [j] 与标准偏差 var _ vector [j]。其中 vector [j] 是特征向量中的第 j 个元素，这样训练数据集中的每个连接数据都可以通过(3)式转换为归一化空间中的新数值。

$$\text{mean\_vector}[j] = \frac{1}{N}\sum_{i=1}^{N}\text{instancei}[j] \tag{1}$$

$$\text{var\_vector}[j] = \frac{1}{N-1}\sum_{i=1}^{N}(\text{instancei}[j] - \text{mean\_vector}[j]^2)^{1/2} \tag{2}$$

$$\text{new\_instance}[j] = \frac{\text{instancee}[j] - \text{mean\_vector}[j]}{\text{var-vector}[j]} \tag{3}$$

$$\text{dist\_vector}(x_1, x_2) = \sqrt{\sum_{i=1}^{d}(x1\ [\text{i}] - x2\ [\text{i}])^2} \tag{4}$$

通过计算每个特征值与平均值之间的标准偏差，可得到该特征值在归一化空间中的新值。对所有的连接数据都作该操作，就实现了在训练数据集提取出的统计信息的基础上，将连接数据由其初始空间转换到新的归一化空间中去。

用聚类方法进行入侵检测的一个重要前提是，同类型数据，无论同是正常数据或攻击数据，都将在特定的度量空间上聚集在一起。但应该如何确定该特定的度量空间呢？这需要对特定的问题加以考虑，在入侵检测中，某些特征如 num _ failed _ logins 应该比 num _ file _ creation 更能说明异常行为（即潜在的攻击）的出现，则前者的差异在构成特征向量间距离方面应比后者更重要。因此，要确定一个最合适的度量，必须尝试不同权值分配的特征子集组合。但实际上，既要对不同特征之间的关系进行定性分析，又要确定其定量比例，将给我们的系统实现带来较大的困难。更重要的是，这种对特定领域问题、特定数据分布以及特定特征集经过精心调整得到的参数必将损害该入侵检测系统的普遍性。正是基于上面的考虑，同时由于已经将数据集转换到一个标准的单位空间中，因此，在我们的系统中，还是采用了标准的欧几里得（Euclidean）距离度量来计算特征向量间的距离，见（4）式。

1.2 聚类算法

在完成了训练数据集的归一化后，我们需要对其进行搜索，并建立一个与其中数据分布相适应的类的集合，并且将每个数据项划分到各自的类中。具体的聚类过程如下所示。

第一步，划分样本子集。因为基于网络的一些攻击行为在使用协议及服务类型上往往有着共性，故而可以据此先将网络连接数据划分成各个不相交的子集，并在子集的基础上再进行类别搜索。

第一级划分，以连接数据中第 7 位（land 位）为依据，因为由该位可直接判断是否为 land 攻击。第二级划分以第二位协议类别为依据，因为很多攻击类型都是分别基于同一种协议的，如 smurf，teardrop。对那些同时以多种协议类型体现的攻击，也可将属于不同协议的连接作为其一个子集。由于我们的根本目的就是要区分正常与异常，所以即使同一种类型的攻击被划分到不同类中去，也是可以接受的。事实上，这样的划分对于我们将少量的攻击数据从正常行为数据中区分出来也是有好处的。第三级的划分可根据服务类别进行。

第二步，寻找类质心集合。依次计算样本子集中数据与类集合中已建立的各类质心的距离，这是决定该数据类划分的重要指标。虽然我们把所有的样本数据集的特征向量都转换到了同一个标准的单位空间，但在类的搜索与建立阶段，为了减少计算时间，我们针对已划分的每个样本子集进行。

在该算法中，类集合的搜索与建立是动态更新的。首先，初始化一个空的类集合。在归一化训练数据集中取出的第一条连接数据自然就作为第一个类的质心保存在类集合里，然后每取出一个新的连接数据 s'，都计算其与类集合已存在的各个类质心的距离。取 $\min\_\text{distance} = \min\{\text{dist}_i(s', C_i), \text{for } i \in 1, \cdots, d\}$。$\text{dist}_i(s', C_i)$ 为 s' 与类集合中 C_i 类的距离。d 为当前类集合中已建立的类的个数，它随着算法的不断进行而逐渐增加。若该 $\min\_\text{distance}$ 值小于或等于我们事先定义的类宽度值 W，则继续处理下一条数据。如果当前数据与类集合中所有类的质心间的距离都大于 W，则认为在类集合中不存在与它相对应的类。此时便以该数据为质心定义一个新的类，并加入到类集合中。当训练样本集中的所有数据都处理完后，算法就进入数据的类划分阶段。

第三步，数据的类划分。此时，系统已得到一个数目确定的类的集合。通过分别计算训练样本集留下的数据与类集合中所有类质心的距离，取出其中的 min _ distance，并将该数据划分到该 min _ distance 对应的类中。

第四步，类质心的调整。由于类质心的选择完全是随机的，我们不能肯定目前处于类质心上的数据能最好地反映数据的分布。因此，须再对每个类计算其质心，即计算各个类中所有数据的均值特征向量，并以该结果作为该类的质心。虽然质心的调整不会影响训练数据集中的数据类划分，但更合理的类分布将更好地用于后续的数据测试及实时检测中。

第五步，区分正常与异常类。到目前为止的工作仅仅是通过聚类算法将相似的数据组合起来，但在所得到的这许多类中，到底哪些是正常连接数据，哪些代表了入侵连接数据，还需要做进一步的分析。本系统设计的基本思想就是从可能包含未知的、新的入侵的训练数据中，建立类划分，并发现那些新的攻击行为。显然，我们无法事先对该类攻击行为进行特征提取、编码，基于滥用检测的方法是无效的。此时，我们能用于进行入侵检测的一个重要假设就是，在采集的训练数据集中，正常行为的类及其子类数据在数量上将远远大于各种体现攻击行为的类及其子类数据。这样，聚类算法所得到的类划分就能够帮助我们区分正常与异常类。一种作法就是将所有的类按其中包含的数据量大小排列，并设定一个比例数 N，那些位于 N 以上的包含最多数据量的类被判断为正常类，而其余的类则被认为是异常类。这种方法非常简单，且易于理解并实现，但其有效性与正常行为的子类数目有密切的关系。如果正常的行为类被划分过细，每个子类都在特征空间中有其独特的类质心，则必将导致单个子类中的数据量相对减少，甚至少到小于某些异常类包含的数据量。在这种情况下，就会错误地将正常的数据类划分为异常，或者是将异常的类划分为正常。为防止该问题的出现，应在生成训练数据集时尽量增大各类正常数据的容量，使得在任何子类中，都能包含足够多的数量以从异常类中区分出来。

1.3 实时检测入侵

一旦从训练数据中得到分类器，就可将之用于入侵的实时检测中。因为这些类的个数是非常有限的，需要的计算量相当小，足以满足实时检测的要求。对我们的系统而言，检测入侵其实也就是数据分类的过程。给定一个连接数据 s，首先需要进行归一化，将它转换到训练数据集所在的特征度量空

间中。在转换过程中所用到的统计信息，如特征向量的均值和标准偏差，依然采用的是在训练数据集中的现有参数。令 s ' 为得到的新特征向量。然后计算 s ' 与类集合中所有类质心的距离 $dist_i$ （s ', C_i），$i \in 1 \ldots d$，C_i 为类集合中的类，d 为类的总数。找到 K 个与 s ' 距离最小的类，如果在这 K 个类中，有半数以上的类为正常类，则判断该连接数据为正常连接，否则判断其为异常。为避免得到的正常类数目与异常类数目相同，K 值一般取为奇数常量。当然在实时检测过程中，也可以只选择最近邻的一个类，并将当前数据划分到该类中。但考虑到数据类分布可能存在的不准确性以及训练数据集与测试数据集之间数据分布可能存在的偏差，选择最近邻的 K 个类，能在统计意义上尽量减少这些问题造成的不利影响。

2 系统性能分析

2.1 各参数的确定

在进一步讨论系统的性能之前，有必要先确定所用到的三个参数值，即聚类宽度 W，在聚类过程中决定在什么样的距离下，两个连接数据必须被分配到同一个类，也是系统类型数据在特征空间中构成的类的平均半径；正常类比 N，在区分正常与异常类时决定正常类在所有类中所占的比例；近邻个数 K，在实时检测中需要考查的最近邻个数。

解决该类问题的一种常用办法就是试探法（trial _ and _ error method）。为了简化问题，假设 $K = 1$，$W = 20$，先考查 N 的取值情况。实验结果如下表1所示。

表1 固定 K 和 W，仅改变 N 的情况下系统性能

Tab.1 System performance with varied N and fixed K, W

K	W	N（%）	检测率(%)	误报率(%)
1	20	10	82.1	9.2
1	20	15	70.4	5.14
1	20	20	62.3	1.4
1	20	25	46.1	0.52

表2 固定 K 和 N，仅改变 W 的情况下系统性能

Tab.2 System performance with varied W and fixed K, N

K	W	N（%）	检测率(%)	误报率(%)
1	10	20	58.5	1.85
1	20	20	62	1.4
1	30	20	68.9	0.82
1	40	20	60.4	0.9

从表1中可见，检测率与误警率都与 N 的取值成正比关系，这与我们估计的情况是一致的。因为将类集合中的类按其中包含数据量大小进行排列之后，大体上是正常数据在前，异常数据集中在后。这样，N 取值越小，也即判断为异常的数据越多，检测率显然要增加。在理想的情况下，聚类结果形成的每个类都将只包括同种类型的数据，要么是正常数据，要么是入侵数据，但实际情况下，这是不可能的，各个类都不可避免地会含有被错误地分配来的数据。对入侵检测环境而言，也就是在正常类中含有异常数据，异常类中也含有正常数据。因此，随着 N 的变小，检测率的增加，误警率也就自然增加了。综合检测率与误警率的情况，我们选用 $N = 20\%$ 作为系统参数，并在后续的实验中继续使用。

下面讨论 W 的取值情况。通过对实验结果的分析，由表2，可见在 $W = 30$ 时，系统有着最佳的性能表现，此时检测率最大，而误警率最低。最后在 W 与 N 都确定的情况下，讨论最近邻数 K 的取值。结果如表3所示。实验数据表明，系统在 $K = 5$ 时有着较好的性能。

通过以上各组实验分析，最后我们得到 $W = 30$，$N = 20\%$，$K = 5$ 作为系统参数，用于对系统的进一步评价。

2.2 系统性能

图1给出了在选用确定的 W、N、K 值时系统检测率与误警率间的关系。研究检测率与误警率之间的制约关系往往是评价入侵检测系统设计性能的一个重要指标，系统设计总是希望检测率能尽可能高，而误警率又尽可能低。但实际上，只能在二者之间取折衷的选择。图中的结果与我们想象的情况是一致的，检测率的提高自然会伴随着误警率的上升。

表 3 固定 W 和 N，仅改变 K 的情况下系统性能

Tab.3 System performance with varied K and fixed W, N

K	W	N（%）	检测率（%）	误报率（%）
3	30	20	62.4	0.97
5	30	20	67.2	0.8
7	30	20	68.9	1.64
9	30	20	61.7	0.83

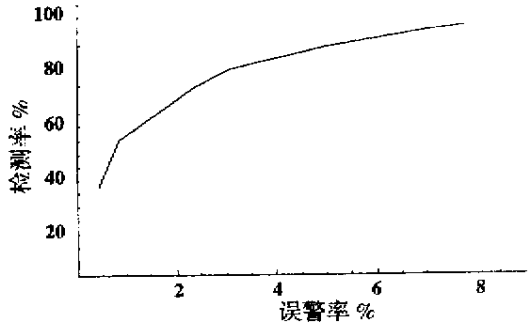

图 1 $W=30$，$N=20\%$，$K=5$ 时的系统性能

Fig.1 System performance with $W=30$, $N=20\%$, $K=5$

2.3 测试结果

尽管在我们的聚类及检测过程中，并没有用到训练数据的类标识信息，系统在训练数据集上的性能可以看作是在测试数据集上的性能，但由于不同数据集的数据分布难免有所差异，在一定的差异范围内，系统能否稳定工作是我们很关心的问题。下面我们通过在测试数据集上的检测，比较在不同数据集上的检测率与误警率。由表 4 结果我们可以作出如下两点判断：一是在数据分布情况上看，训练数据集与测试数据集是基本相同的。二是系统在聚类与检测上的性能是稳定的。

表 4 系统在训练数据集与测试数据集上的性能比较

Tab.4 System performance comparison between training dataset and test dataset

	检测率%	误报率%
训练数据集	67.2	0.8
测试数据集	65.8	0.75

3 结 论

通过对常见的入侵检测方法的分析，指出了其中存在的缺陷。尤其是在训练数据集的收集问题上，当前很多用于入侵检测的机器学习算法都有着相当严苛的要求，或者是需要完全洁净的正常数据，或者是需要对所有样本数据作出正确的标识。针对上述情况，本文提出了一种基于聚类的入侵检测方法，该方法只需在训练数据集中各类正常数据的数量都远远大于所有异常数据类。实验表明，与传统的方法相比，基于聚类的方法尽管在检测性能上有一些差距，但由于其应用的便利及对训练数据集的低要求，该方法在网络入侵检测领域还是有着广泛的应用前景。

在 W、N 和 K 的确定上，该方法还存在着较大的随意性，下一步的工作将继续研究如何对上述各值进行确定的问题。

参考文献：

[1] Eskin E. Anomaly detection over noisy data using learned probability distributions [A]. Proceedings of the International Conference on Machine Learning, 2000.

[2] Axelsson Stefan. Intrusion Detection Systems: A Survey and Taxonomy [EB]. http://citeseer.nj.nec.com/sc, 2000.

[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).

[4] Bonifacio J M, Cansian A M. Neural Networks Applied in Intrusion Detection Systems [EB]. http://citeseer.nj.nec.com/sc, 2000.

[5] Lee W, Stolfo S J. Data Mining approaches for intrusion detection [A]. Proceedings of the 1998 USENIX Security Symposium, 1998.

[6] KDD Cup 1999 Data [EB]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. 1999.

[7] 张平安，高春华等译. 神经—模糊和软计算 [M]. 西安：西安交通大学出版社，2000.