第37卷 第17期
Vol.37 No.17

计 算 机 工 程
Computer Engineering

2011年9月
September 2011

• 安全技术 •

文章编号：1000—3428(2011)17—0096—03 文献标识码：A 中图分类号：TP309; TP393

基于聚类分析的网络入侵检测模型

李文华

(长江大学计算机科学学院，湖北 荆州 434023)

摘 要：为提高网络入侵检测系统的入侵识别能力，提出一种基于模糊C均值(FCM)聚类的入侵检测模型。该模型包括数据预处理器、FCM聚类处理器、类中心集更新器和检测系统，可以同时处理数值属性与符号属性。实验结果表明，与其他模型相比，该模型具有较低的误警率和较高的检测率。

关键词：入侵检测；聚类分析；模糊C均值；欧氏距离；简单匹配系数

Network Intrusion Detection Model Based on Clustering Analysis

LI Wen-hua

(College of Computer Science, Yangtze University, Jingzhou 434023, China)

【Abstract】This paper introduces Fuzzy C-means(FCM) clustering method, researches the methods of intrusion detection based on clustering analysis, and establishes a new model of network intrusion detection. The new model is included data pre-processor, clustering-component based FCM, Updater of clustering-center, and detection system, and improves the availability of intrusion detection system. Experimental result proves that the model can detect intrusion from the network connection data at a lower system false alarm rate and a higher detection rate.

【Key words】intrusion detection; clustering analysis; Fuzzy C-means(FCM); Euclidean distance; Simple Matching Coefficient(SMC)

DOI: 10.3969/j.issn.1000-3428.2011.17.031

1 概述

网络入侵检测系统(Intrusion Detection System, IDS)是在防火墙和信息加密等技术之后发展起来的新一代网络安全防护系统，是一种能够动态监控和防御网络系统入侵行为的安全机制。其过程是通过网络中的若干关键点收集并分析数据包，从中发现这些包中是否包含具有异常行为特性的数据并做出相应的报警响应[1]。根据所采用的数据分析技术不同，入侵检测可以分为误用检测和异常检测2类[2]。聚类分析是数据挖掘领域中最成功的一种利用无监督学习过程获取知识的方法。本文在研究基于聚类分析的入侵检测方法基础上，建立一种基于模糊C均值(Fuzzy C-means, FCM)聚类的网络入侵检测模型。

2 基于聚类的网络入侵检测方法

基于聚类的入侵检测方法是在选定的训练数据集上进行聚类分析并以此建立行为模型，然后利用此模型评估网络实时数据来检测入侵。文献[1]实现了基于聚类分析的入侵检测模型，其检测率可达到50%，误警率在1%左右。该模型的检测过程如下：

(1)对训练数据进行预处理，把数值属性进行标准化处理，并对符号属性进行编码，把符号属性转化为数值属性。

(2)对预处理后的数据进行聚类分析，得到聚类中心和各个类的实例数。聚类算法选择简单且易于实现的单链聚类法。

(3)把类中实例数少于某个设定阈值的聚类中心作为异常类中心，超过阈值的聚类中心作为正常类中心。

(4)对每一条预处理过的实时连接，计算其与各个类中心的距离，如果与异常类中心最近，则认定该连接为异常，并进行报警处理；否则，认定该连接属正常网络请求。

该模型的提出基于2个假设。假设1：网络中正常行为与入侵行为有明显的不同，因此，在对特征空间和度量单位的合理选择下，可以通过聚类分析方法将正常和异常行为加以区分，并聚集到不同的簇中；假设2：网络中正常行为数据远远超过入侵行为数据。这样通过简单的阈值设定，就可以把那些实例数少于该阈值的类作为异常类。

文献[2]在上述模型基础上进行改进，通过引入K-means聚类算法增强了这种无监督的入侵检测方法，实验结果表明，在聚类半径 L=40，正常类比 N=20%时该算法的整体性能较好，其检测率可达到60%以上，误警率保持在1%以下。

分析上述无监督入侵检测模型，可以发现以下不足：

(1)模型中的数据预处理没有对属性进行选择，而在实际获取的网络数据中，属性数往往比较大，且不是所有的属性都对入侵检测有作用，庞大的属性数影响了入侵检测模型的性能。

(2)模型在训练阶段，采用的单链法聚类或K-means聚类等均属硬性聚类，虽然具有一定的高效性，但不能很好地体现数据集性态和类属方面的中介性，所形成的用于检测入侵的聚类中心与数据集的最优中心相差较大。

(3)对于符号属性，这类模型考虑在数据预处理阶段通过编码的方式把符号属性转化为数值属性，具有一定的指导意义，但是在对模型进行测试时，某个符号属性的全部离散值很难获取，另外随着计算机网络传输的发展，这些符号本身也是不断扩展和增加的，经常会出现一些无法识别的符号，这时模型就会产生错误。

(4)上述入侵检测均属于无监督检测方法，这类方法必须基于假设2，只有在此假设基础上，才可以根据聚类的大小来判断是否为入侵数据。但此假设并不总是适用，有些入侵

作者简介：李文华(1965－)，男，副教授，主研方向：网络安全，数据库技术

收稿日期：2011-03-04 **E-mail：**wenhua999@qq.com

如 DoS，经常会产生大量的入侵数据，而某些类型的正常系统行为却只产生少量的数据。显然，在这 2 种情况下无法获得良好的检测效果。

3　基于 FCM 的网络入侵检测模型

为弥补上述入侵方法的不足，本文提出一种基于 FCM 的网络入侵检测模型。该模型如图 1 所示。

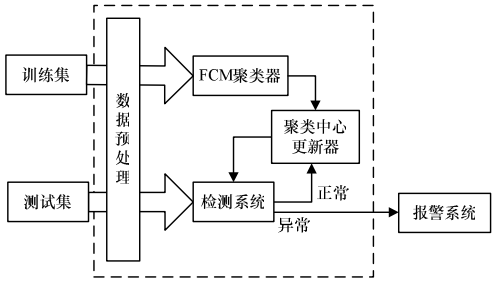

图 1　基于聚类分析的网络入侵检测模型

从图 1 可以看出，本文提出的模型主要由 4 个部件组成：数据预处理器，FCM 聚类处理器，类中心集更新器和检测系统。下面分析各个组件的作用和实现方法。

3.1　数据预处理器

要提高入侵检测系统的性能，在对网络数据进行分析之前必须进行预处理。网络数据预处理包含很多内容，本文在检验模型时使用的实验数据选自 KDDCUP99 数据集，相对比较“干净”，所以，数据预处理器只对网络数据进行属性选择和标准化处理。

针对属性选择，本文直接使用文献[3]的研究成果，该文献利用支持向量机计算出在 KDDCUP99 数据集中有 13 个属性最为重要，其中数值属性 11 个，符号属性 2 个。

数据标准化的处理对象是数值属性。这是因为对于数值属性来讲，不同的属性特征有不同的度量标准，如果不进行标准化处理，就会出现大数吃小数的情况，造成计算偏差。数据标准化又称为数据归一化，其过程如下：

令 $\overline{X}_j$，R_j 和 S_j 分别表示第 j 个属性的样本均值、样本极差和样本标准差：

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{1}$$

$$R_j = \max_{1\leqslant i\leqslant n}\{x_{ij}\} - \min_{1\leqslant i\leqslant n}\{x_{ij}\} \tag{2}$$

$$S_j = [\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{X}_j)^2]^{1/2} \tag{3}$$

则标准化的数据为：

$$x'_{ij} = \frac{x_{ij} - \overline{X}_j}{S_j} \tag{4}$$

x'_{ij} 即为标准化后的实例数据值。经过这种标准化之后，各个数值属性的极大值为 1，极小值为 0。相当于利用统计特性将原始数据的属性值映射到一个标准的属性空间，以便于减少上述偏差问题。

3.2　聚类分析器

聚类分析器用于对训练集进行聚类分析以生成初始的类中心集，属核心部件。本文在实现模型时，使用 FCM 作为部件的核心算法。FCM 算法[4]是一种基于模糊划分的聚类算法，它把 n 个向量 x_i 分为 c 个模糊的组，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。它以所分组内距离的平方和最小化为依据，用隶属度确定每个数据样本属于某个聚类的程度。

模糊 C 均值的价值函数表示为：

$$J(\boldsymbol{U}, c_1, c_2, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2(X_i, C_j) \tag{5}$$

$$\text{s.t.}\quad \sum_{j=1}^{k} u_{ij} = 1, \forall i = 1,2,\cdots,n \tag{6}$$

其中，$\boldsymbol{U} = (u_{ij} \mid i = 1,2,\cdots,n, j = 1,2,\cdots,k)$ 为分类矩阵，元素 u_{ij} 表示第 i 个数据样本属于第 j 类的隶属度；$d_{ij}(X_i, C_j)$ 为第 i 个数据样本与第 j 个聚类中心之间的距离；参数 $m > 1$ 为模糊系数，也叫加权指数，用来控制分类矩阵 $\boldsymbol{U}$ 的模糊程度，m 越大越模糊；式(6)是价值函数的约束条件，指一个数据集的隶属度和总要等于 1。

应用拉格朗日乘数法，使式(5)得到最小值的必要条件为：

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{7}$$

$$u_{ij} = \left(\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}\right)^{-1} \tag{8}$$

通过对式(7)和式(8)进行迭代运算，即可求出分类矩阵和聚类中心集，且由于 $m > 1$，因此该运算是收敛的。分类矩阵表示的是每个样本点属于每个类的隶属度。根据此矩阵按模糊集合中的最大隶属度原则，就能确定每个网络连接归为哪类。

对于聚类分析中使用的距离度量 d 的计算，通常使用欧式距离，但该距离度量存在一个缺陷，即只能用于对数值属性的计算。对于符号属性，在以往基于聚类的入侵检测中都通过编码的方式把符号属性转化为数值属性，这样的方法在检测新加入的入侵数据时可能会产生无法识别的错误，所以，引入简单匹配系数(Simple Matching Coefficient, SMC)来对符号属性进行单独处理[5]。下面分别从形式化角度定义数值属性与符号属性的距离度量公式：

对数值属性，距离度量使用的是欧氏距离，计算公式为：

$$d(x,y) = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2} \tag{9}$$

其中，x、y 代表 2 个实例；n 表示数值属性集的个数。

对于符号属性，用 SMC 来进行度量，其度量公式为：

$$d(x,y) = \frac{\sum_{k=1}^{p} g(x_k, y_k)}{p} \tag{10}$$

其中，$g(x_k, y_k) = \begin{cases} 0, x_k = y_k \\ 1, x_k \neq y_k \end{cases}$；$x$、$y$ 分别代表 2 个数据实例；p 表示符号属性集中的元素个数。对于聚类中心中该符号属性的设定，取该类中所有实例属性值最多的那个类值。

经过聚类分析器的处理，生成的正常类中心集 $NC = \{nc_1, nc_2, \cdots, nc_n\}$ 和异常类中心集 $AC = \{ac_1, ac_2, \cdots, ac_m\}$，并记录每个符号属性内所有出现的值及对应在训练集中出现的实例个数。

3.3　检测系统

该部件主要用于对数据和类中心之间的距离进行度量，并根据度量结果对连接进行类属标示。在模型完成训练后，进入测试阶段，检测系统根据类中心更新器提供的类中心集，对每条用于测试的连接与两个类中心集元素进行距离度量，距离度量计算公式参照式(9)和式(10)。依据距离最小的那个类中心的性质，对连接进行标记。把标记后的数据送入报警系统，报警系统根据连接的标示，作出相应的反应。

此外，为使类中心集能随着网络环境的变化作出响应的更新，对每条标记的连接，送入类中心更新器，让类中心更新器根据连接对类中心集中对应的元素进行更新处理。

3.4 类中心更新器

类中心更新器在训练阶段用来保存由聚类分析器生成的类中心集；在测试阶段，利用检测系统对一条连接检测并标记后，类中心更新器要根据检测系统对连接的标记结果，对相应类的中心进行更新处理，步骤如下：假设已测得网络连接 $x=\{x_1,x_2,\cdots,x_n\}$ 与某一类中心 $c=\{c_1,c_2,\cdots,c_n\}$ 距离最近，则检测系统就把该连接标记为类 c。

根据数据所标示的类属，对类中心 c 作如下更新处理：对于数值属性，把连接和中心对应的属性值相加取均；对于符号属性，把更新器中记录的与该连接相同的那个符号值的实例数加 1，比较所有符号值出现的实例个数，选取最多的那个作为类中心在该属性上的值。

4 仿真实验

4.1 数据描述及选取

为评价本文模型的有效性和可行性，选用 KDDCUP99 网络数据包[6]进行实验分析。该数据包由麻省理工学院 Lincoln 实验室从一个模拟美国空军局域网上采集，已成为很多基于网络入侵检测常用数据包，对验证 IDS 性能具有较好的指导性意义。整个基本数据包中有 2 个数据集：用于训练模型的训练集和用于测试模型的测试集。2 个数据集的具体数据信息如表 1 和表 2 所示。可以看出已知的攻击有 4 种，在测试集中，有一部分未知的攻击，用于测试模型对未知攻击的检测性能。

表 1 训练集数据信息

正常实例数	攻击实例数					总实例数
	DoS	U2R	R2L	Probing	共计	
97 278	391 458	52	1 126	4 107	396 743	494 021

表 2 测试集数据信息

正常实例数	攻击实例数						总实例数
	DoS	U2R	R2L	Probing	未知攻击	共计	
60 593	223 298	39	5 993	2 377	18 729	250 436	494 021

仿真实验首先从 KDDCUP99 训练集中抽取用于训练模型的 2 组训练数据集：一组为只包含正常实例的数据集，其实例数为 2 000 条；一组为只包含异常实例的数据集，其实例数为 8 000 条。在选择异常实例时，根据 4 种攻击类型的比例进行选取：DoS 攻击类型选取 7 200 条，U2R 攻击类型选取 50 条，R2L 攻击类型选取 250 条，Probing 攻击类型选取 500 条。此外，从测试集中随机抽取 4 组 2 000 个样本的数据集。表 3 为所抽取的测试数据集攻击类型分布信息。

表 3 数据集攻击类型分布信息

数据集	正常实例数	入侵实例数	入侵实例分布				
			DoS	U2R	R2L	Probing	未知攻击
1	483	1 517	993	4	57	104	342
2	507	1 493	1 017	3	53	99	321
3	492	1 508	1 008	5	51	116	327
4	511	1 489	996	0	62	93	338

4.2 实验结果分析

利用 Java 和 Matlab 共同实现本文提出的模型，程序运行系统环境 Windows XP，硬件环境 Intel Core 2 Duo CUP(时钟频率 2.00 GHz)，内存 1 GB。

聚类半径是实验需要提前输入的参数，经过多试探性实验，确定该参数为 10。实验中检测率定义为：算法正确检测到入侵个数与测试集所有入侵个数的百分比；误警率定义为：模型中把正常数据错误检测为异常数据的个数与所有正常数据个数的百分比。

实验首先把经过预处理的 2 个训练集送到 FCM 聚类器中进行聚类分析，生成各自的聚类中心；然后把选出的 4 个测试集分别进行相应的预处理并送入检测系统中进行入侵检测。检测结束后，根据标示的类属，计算出针对每个数据集的检测率和误警率。

图 2 给出了 4 个测试数据集的平均检测率和误警率的 ROC(Receiver Operating Characteristic)曲线。从中可知，最好性能是当检测率为 85%时，误警率为 1.5%。

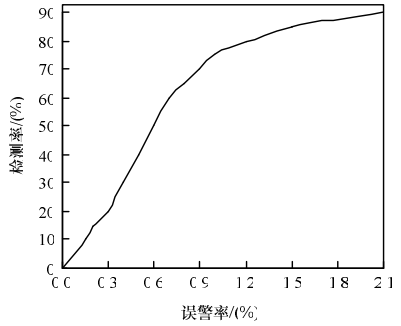

图 2 检测率和误警率的 ROC 曲线

可以看出，本文模型的入侵检测性能要优于文献[1-2]的模型。另外，在实验中增加了对未知入侵的检测，从相对曲线来看，未知入侵的加入并没有影响本文模型的性能，说明本文提出的模型对未知入侵是健壮的。

5 结束语

目前网络安全技术主要包括加密和数字签名技术、身份认证与访问控制技术、防火墙技术、入侵检测技术和入侵诱骗技术。这些安全机制相互协作，在为营造一个安全的网络环境方面提供了强大的技术保障。本文分析现有基于聚类分析的网络入侵检测方法，设计并实现一种网络入侵检测模型。该模型改进了已有模型的不足，增强了高效性和健壮性 2 个方面的性能，实验结果证明了本文模型可取得较高的检测率与较低的误警率，是一种有效的入侵检测方法。本文在数据预处理中利用了有监督的处理方法，下一步将对无监督的处理方法进行研究，以扩大该模型的应用范围。

参考文献

[1] Portnoy L, Eskin E, Stolfo S J. Intrusion Detection with Unlabeled Data Using Clustering[C]//Proc. of ACM CSS Workshop on Data Mining Applied to Security. Philadelphia, USA: ACM Press, 2001.

[2] 罗 敏, 王丽娜, 张焕国. 基于无监督类的入侵检测方法[J]. 电子学报, 2003, 31(11): 1714-1716.

[3] Mukkamala S, Janoski G, Sung A H. Intrusion Detection Using Neural Networks and Support Vector[C]//Proc. of IEEE Int'l Joint Conference on Neural Networks. Honolulu, Hawaii, USA: [s. n.], 2002.

[4] 吴 静, 刘衍珩, 吕 荣. 基于 FCM 的分布式学习方法[J]. 吉林大学学报: 工学版, 2010, 40(1): 171-175.

[5] 任 晓, 张永奎, 薛晓飞. 基于 K-modes 聚类的自适应话题追踪技术[J]. 计算机工程, 2009, 35(9): 222-224.

[6] MIT Lincoln Lab.. KDDCUP99 Dataset[DB/OL]. [2010-05-11]. http://kdd.ics.uci.edu/databases/kddcup99.

编辑 金胡考