ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.8, August 2008, pp.2140−2148
DOI: 10.3724/SP.J.1001.2008.02140

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

网络入侵检测中的自动决定聚类数算法*

肖立中[1+], 邵志清[2], 马汉华[2], 王秀英[2,3], 刘 刚[2]

[1](上海应用技术学院 计算机科学与信息工程系,上海 200235)

[2](华东理工大学 信息科学与工程学院,上海 200237)

[3](上海新侨职业技术学院,上海 200237)

An Algorithm for Automatic Clustering Number Determination in Networks Intrusion Detection

XIAO Li-Zhong[1+], SHAO Zhi-Qing[2], MA Han-Hua[2], WANG Xiu-Ying[2,3], LIU Gang[2]

[1](Department of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai 200235, China)

[2](School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

[3](Xin Qiao Polytechnic Institute, Shanghai 200237, China)

+ Corresponding author: E-mail: lyexp@163.com

Xiao LZ, Shao ZQ, Ma HH, Wang XY, Liu G. An algorithm for automatic clustering number determination in networks intrusion detection. *Journal of Software*, 2008,19(8):2140−2148. http://www.jos.org.cn/1000-9825/19/2140.htm

Abstract: To address the issue in fuzzy C-means algorithm (FCM) that clustering number has to be pre-defined, a clustering algorithm, F-CMSVM (fuzzy C-means and support vector machine algorithm), is proposed for automatic clustering number determination. Above all, the data set is classified into two clusters by FCM. Then, support vector machine (SVM) with a fuzzy membership function is used to testify whether the data set can be classified further. Finally, the result of clusters can be obtained by repeating the computation process. Because affiliating matrix, obtained by the introduction of SVM into FCM, is defined to be the fuzzy membership function, each different input data sample can have different penalty value, and the separating hyper-plane is optimized. F-CMSVM is an unsupervised algorithm in which it is neither needed to label training data set nor specify clustering number. As shown from our simulation experiment over networks connection records from KDD CUP 1999 data set, F-CMSVM has efficient performance in clustering number optimization and intrusion detection.

Key words: fuzzy C-means algorithm; support vector machine; fuzzy membership function; clustering number; intrusion detection

摘 要: 针对模糊 C 均值算法(fuzzy C-means algorithm,简称 FCM)在入侵检测中需要预先指定聚类数的问题,提

* Supported by the National Natural Science Foundation of China under Grant No.60373075 (国家自然科学基金); the Shanghai Education Commission Foundation for Excellent Young High Education Teacher of China under Grant No.YYY-07008 (上海高校选拔培养优秀青年教师科研专项基金); the Open Research Foundation of Shanghai Institute of Technology of China under Grant No.YJ2007-24 (上海应用技术学院引进人才科研启动项目)

Received 2006-07-13; Accepted 2007-05-24

出了一种自动决定聚类数算法(fuzzy C-means and support vector machine algorithm,简称 F-CMSVM).它首先用模糊 C 均值算法把目标数据集分为两类,然后使用带有模糊成员函数的支持向量机(support vector machine,简称 SVM)算法对结果进行评估以确定目标数据集是否可分,再迭代计算,最终得到聚类结果.支持向量机算法引入模糊 C 均值算法得出的隶属矩阵作为模糊成员函数,使得不同的输入样本可以得到不同的惩罚值,从而得到最优的分类超平面.该算法既不需要对训练数据集进行标记,也不需要指定聚类数,因此是一种真正的无监督算法.在对 KDD CUP 1999 数据集的仿真实验结果表明,该算法不仅能够得到最佳聚类数,而且对入侵有较好的检测效果.

关键词: 模糊 C 均值算法;支持向量机;模糊成员函数;聚类数;入侵检测

中图法分类号: TP393 **文献标识码:** A

随着网络技术的不断发展和网络规模的不断扩大,网络入侵的机会也越来越多,网络安全已经成为一个全球性的重要问题.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.

入侵检测技术主要分为两类,即误用检测(misuse detection)和异常检测(abnormal detection).早期入侵检测技术的研究主要集中在误用检测,它依赖于对训练数据集中标记数据样本的学习,当遇到未知攻击时需要用新的标记数据样本对检测系统重新进行训练.然而,标记大量的网络数据代价是很高的.而异常检测可以不依赖标记数据样本而对入侵进行有效检测.聚类检测是一种异常检测技术,它将相似的数据划分到同一个聚类中,而将不相似的数据划分到不同的聚类中,能够自动地对未知攻击进行检测.

模糊 C 均值算法(fuzzy C-means algorithm,简称 FCM)是一种有效的聚类算法,该方法要求预先给定聚类数 k,但聚类数 k 在入侵检测中是不能预先知道的.因此,如何确定聚类数、真正实现无监督的聚类成为一个重要的研究课题.目前,许多学者提出了多种聚类准则[1,2],他们根据不同的聚类准则采用自组织迭代技术或者遗传算法得到最优聚类.但是,采用什么样的适应度函数取决于数据库的结构,使用前需要验证.

最近,模糊支持向量机(fuzzy support vector machine algorithm,简称 FSVM)已运用到入侵检测中来[3].在本文中,我们提出了一种自动决定聚类数算法(fuzzy C-means and support vector machine algorithm,简称 F-CMSVM).它不是直接用模糊支持向量机进行分类,而是用它来帮助模糊 C 均值算法确定聚类数,并且只需进行 2-类支持向量机分类,避免了多类分类的复杂.同时,该算法在确定聚类数的过程中不需要使用适应度函数,避免了适应度函数的选择和验证.

本文第 1 节将介绍这种算法.第 2 节介绍基于这种算法的入侵检测系统.第 3 节介绍实验及其结果.最后在第 4 节给出结论.

1 自动决定聚类数算法

1.1 模糊C均值算法

模糊 C 均值算法是一种基于划分的聚类算法,是普通 C 均值算法的改进.它以所分组内距离的平方和最小化为判据,用隶属度确定每个数据样本属于某个聚类的程度,把 n 个数据样本 $X=\{X_i|X_i\in R^D(i=1,2,\ldots,n)\}$ 分为 k 个类,并求每组的聚类中心 $C=\{C_j|C_j\in R^D(j=1,2,\ldots,k)\}$,使得非相似性指标的价值函数取得最小值.

价值函数表示为

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

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

其中,$U=(u_{ij}|i=1,2,\ldots,n,j=1,2,\ldots,k)$ 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 个聚类中心之间的欧氏距离;参数 $m>1$ 为模糊系数,用来控制分类矩阵 U 的模糊程度,m 越大

越模糊.应用 Lagrangian 乘数法,使式(1)得到最小值的必要条件为

$$C_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}X_i}{\sum_{i=1}^{n}u_{ij}^{m}} \tag{3}$$

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

通过对式(3)和式(4)进行迭代运算,由于 $m>1$,因此该运算是收敛的.

1.2 标准的2-类支持向量机算法

针对一组有标记的样本$(X_1,y_1),(X_2,y_2),\ldots,(X_n,y_n)$,其中,$X_i\in R^D$ 属于两类中的一类,$y_i\in\{-1,1\}$为类别标记.支持向量机算法的主要目的是构造一个分类超平面以分割两类不同的样本,使得分类间隔最大.由此产生以下优化问题:

$$\min\left(\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}\xi_i\right) \tag{5}$$

$$\text{s.t. } y_i(w\cdot X_i-b)\geq 1-\xi_i,\ i=1,2,\ldots,n,$$

$$\xi_i\geq 0.$$

其中,c 为惩罚参数,ξ为松弛变量.将上述问题表示成 Lagrangian 乘子式:

$$\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(X_i,X_j) \tag{6}$$

$$\text{s.t. } 0\leq\alpha_i\leq c,$$

$$\sum_{i=1}^{n}\alpha_i y_i=0.$$

其中,$K(X_i,X_j)$为核函数.本文中使用了径向基核函数(radial basis kernel function,简称 RBF):

$$K(X_i,X_j)=\exp\left\{-\frac{\|X_i-X_j\|^2}{\sigma^2}\right\}.$$

在实验中,我们用 libstvn 自带的 easy.py 脚本调整出最好的 c 和σ.

可以得到分类函数:

$$f(x)=sign\left\{\sum_{i=1}^{n}\alpha_i y_i K(x,x_i)+b\right\} \tag{7}$$

式(5)的第 2 项是惩罚项,较大的 c 意味着为错误项指定了较大的惩罚值,从而减少了错误分类的数据点;较小的 c 意味着忽略了一些“微不足道”的错误分类点,因而可以得到较大的分类间隔(margin).无论 c 的值是大还是小,在支持向量机算法的训练过程中,这个值始终是固定的,这样就导致了算法对某些特殊情形的过分敏感,例如孤立点与噪声,这种情形即是所谓的“过学习(overfitting)”现象.

1.3 引入模糊函数的2-类支持向量机算法

为了防止出现“过学习”现象,文中的训练点不是严格属于两类中的某一类,而是存在一个模糊值 u_i,以区别不同的点对分类结果的影响.这样,样本集变为$(X_1,y_1,u_1),(X_2,y_2,u_2),\ldots,(X_n,y_n,u_n)$.为了解决优化问题,模糊支持向量机算法的 Lagrangian 式为

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}w\cdot w+c\sum_{i=1}^{n}u_i\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(w\cdot z_i+b)-1+\xi_i)-\sum_{i=1}^{n}\beta_i\xi_i \tag{8}$$

其中,$z_i=\varphi(x_i)$表示特征空间中的向量,φ是由输入空间 R^D 到特征空间 Z 的映射.为了找到鞍点,应满足如下条件:

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}\alpha_i y_i z_i=0,$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0,$$

$$\frac{\partial L}{\partial \xi_i}=cu_i-\alpha_i-\beta_i=0.$$

最优超平面的问题就转化为

$$\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{9}$$

$$\text{s.t. } 0\le\alpha_i\le cu_i,$$

$$\sum_{i=1}^{n}\alpha_i y_i=0.$$

最终的分类函数为

$$f(x)=sign\left\{\sum_{i=1}^{n}\alpha_i y_i K(x,x_i)+b\right\} \tag{10}$$

$$\text{s.t. } 0\le\alpha_i\le cu_i$$

模糊支持向量机算法仍采用分类间隔的最大化以及分类错误的最小化,使得分类器具有较好的泛化能力.与传统支持向量机算法不同的是,在模糊支持向量机算法中将惩罚项模糊化,以降低错分数据对分类结果的影响.在此,惩罚项是一个模糊成员函数.显然,模糊支持向量机算法与传统支持向量机算法的Lagrange乘子α_i的上界不同.在后者中,α_i的上界为一个常量c;而在前者中,它的上界是一个动态的模糊函数值,数据点X_i属于其类时所对应的模糊成员函数的值越小,α_i所对应的轴心区域就越窄.

1.4 自动决定聚类数算法及分析

在具备了上面的理论基础以后,我们可以执行自动决定聚类数算法.

首先,假设给定数据集可以分为两类(k=2),用模糊C均值算法进行聚类,然后用算法获得的隶属矩阵作为模糊支持向量机算法的模糊参数对数据集进行训练获得支持向量(support vector,简称 SV)和分类超平面.为了验证假设,需要有一个标准.现有如下两种标准:

1) 使用比率#SV/n[4]进行检测.但是文献[5]中指出,在高维数据集中很难获取支持向量的个数.由于 KDD99数据集的维数较高,所以我们不采用这一标准.
2) 在本文提出的新标准中,用 d_{SV}表示两类之间的距离,用 d_{S1}和 d_{S2}表示两类 $S1$ 和 $S2$ 中各自的支持向量与 h 个最近邻点的平均距离.如果 $d_{SV}\le\min(d_{S1},d_{S2})$,则原数据集不可分,假设不成立;否则,假设成立,原数据集至少可以分为两类.

在支持向量机算法中,通常用 $2/\|w\|$表示两类之间的距离,因此 $d_{SV}=2/\|w\|$.当 $d_{SV}\le\min(d_{S1},d_{S2})$时,说明两类之间的边界密度大于或等于两类中某类的密度分布,因此划分不合理,数据集中不存在聚类结构;相反,当 $d_{SV}\ge\min(d_{S1},d_{S2})$时,说明在数据集中至少存在两个区域有明显的界限,因此至少可以分为两类.用此种标准进行判断,运算简单,在后面的实验中也证明它是有效的.

下面是自动决定聚类数算法的描述:

令 $ncluster(Si)$表示数据集 Si 的聚类数,nc 表示输入数据集总的聚类数.

1) 对输入数据集 S,令 k=2,用值在 0,1 间的随机数初始化隶属矩阵 U,并使其满足式(2)的要求;
2) 用式(3)计算聚类中心;
3) 用式(4)计算隶属矩阵.

计算式(1),如果 $J_m(U,C)$小于某一阈值或与上次的改变量小于某一阈值,则算法停止,输出聚类结果 $S1$ 和 $S2$;否则,转步骤 2).

4) 用步骤 3)所得隶属矩阵 U 作为模糊支持向量机算法的模糊参数,以分类结果 $S1$ 和 $S2$ 作为标记对 S 进行模糊支持向量机算法训练;

5) 计算 d_{SV},d_{S1} 和 d_{S2};

6) 如果 $d_{SV}\leq\min(d_{S1},d_{S2})$,则 S 不可分,nc=1 算法结束;否则,S 可以分为 $S1$ 和 $S2$,

$$nc=ncluster(S1)+ncluster(S2);$$

7) 把 $S1$ 和 $S2$ 分别作为 S,转步骤 1).

很多聚类算法都有这样一个缺陷:需要预先指定聚类数.模糊 C 均值算法也是如此.本文之所以采用模糊 C 均值算法进行聚类,是因为它对数据的划分是柔性的模糊划分,相对于普通 C 均值算法的硬性划分可以取得更好的聚类效果.同时,算法得到的隶属矩阵可以运用到模糊支持向量机算法中,以得到更好的分类效果.

在基于聚类算法的网络入侵检测算法中,很多是通过事先设定最大聚类数后再利用穷举法选定最佳聚类数,或是先设定一个聚类数再进行调整,如文献[6,7],这对于维数较高、数据量较大的网络数据是很不方便的,因为网络入侵模式是不断改变的,入侵检测系统需要不断地训练和更新,每次都对大量的网络数据进行分析,以确定聚类数范围再进行调整,需要专家知识和大量的工作.而本文的算法可以自动确定聚类数,不需要人工干预.

为了验证本文算法对确定聚类数的准确性,我们构造了一个平衡类中距与类间距的适应度函数:

$$f(S)=\frac{a}{J_m+\frac{1}{k}d^2(C)} \tag{11}$$

其中,a 为常数,针对不同的数据集而定;J_m 为模糊 C 均值算法的价值函数;$d^2(C)=\sum_{i,j=1}^{k}\|C_i-C_j\|^2$ 为各聚类中心之间欧氏距离的平方和.

在判断聚类数是否为最优时,仅使用类中距 J_m 是不够的.因为当聚类个数 k 增加时,类中距呈递减趋势而类间距 $d(C)$总体呈增加趋势.我们分类的主要目标是使类中距最小、类间距尽可能地大.为了限制类中距单调,我们构造的适应度函数加上了惩罚因子 $d^2(C)$,使得适应度函数能够平衡类中距和类间距,迫使聚类数不能等于样本数 n.设想一下,当聚类个数 k=n 时,类中距虽然可得最小值,但这样的分类是没有意义的.同时,为了区分两项的主次,惩罚因子 $d^2(C)$的权值设为 $k\geq2$.

2 基于自动决定聚类数算法的入侵检测系统

基于自动决定聚类数算法的异常入侵检测系统主要由数据预处理器、自动决定聚类数分类器和检测系统三部分组成,如图 1 所示.

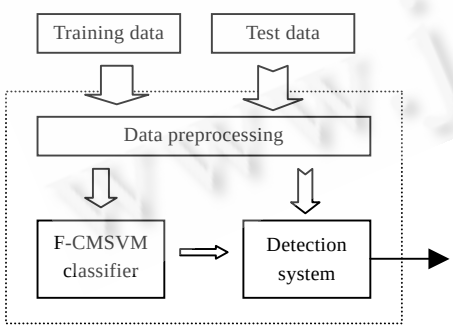

Fig.1 The intrusion detection system based on F-CMSVM

图 1 基于自动决定聚类数算法的入侵检测系统

数据预处理器用来对输入数据进行属性的选择和预处理;自动决定聚类数分类器用来对训练数据进行聚类,并把相应的聚类中心提供给检测系统;检测系统用来对测试数据作出判断,判定它是正常还是入侵.

基于无监督聚类的入侵检测算法往往建立在一个假设之上,即正常行为的数目远远大于入侵行为的数目.在这个假设成立的前提下,入侵检测可以根据聚类的大小来判断.大的聚类对应正常数据,小的聚类往往就是入侵.然而,有些入侵例如 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.这两种情况都不满足以上假设,往往需要单独处理[8].在本文中,我们首先采用只包含正常数据的数据集作为训练数据,再针对包含入侵数据的数据集进行检测,计算出测试数据与各聚类中心的距离 d,当 d 小于某一设定的阈值时,该数据被认为是正常数据;否则被认为是入侵数据.由于我们的检测不是根据聚类的大小来进行的,所以在以上

假设不成立的情况下也能进行检测.

整个系统的工作过程分为两个阶段:训练阶段和检测阶段.在训练阶段,自动决定聚类数分类器对正常数据进行处理,得出聚类划分和聚类中心;在检测阶段,检测系统根据分类器提供的聚类中心和设定的阈值对测试数据进行判断.

3 实验与实验结果

3.1 实验数据描述

为了评价本文算法在入侵检测中的效果,我们选用 KDD Cup 1999 网络数据集进行实验.该数据集是麻省理工学院 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.数据集中包括多种网络环境下的模拟入侵,其中每个连接实例包含 41 个属性.

为了进行仿真实验,我们首先从 KDD Cup 1999 数据集中随机抽取一部分正常数据作为训练数据,然后再从中随机抽取 3 个样本数据集 *D*1,*D*2 和 *D*3,各包括 3 150 个数据作为测试数据.KDD Cup 1999 数据集中的入侵类型按攻击手段可大致分为 Probing,DoS,U2R,R2L 四类.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 3 个数据集,使得每个数据集中有 3 000 个正常连接、150 个异常连接,而且使异常连接的种类尽可能平均.各数据集中攻击类型的分布如图 2 所示.

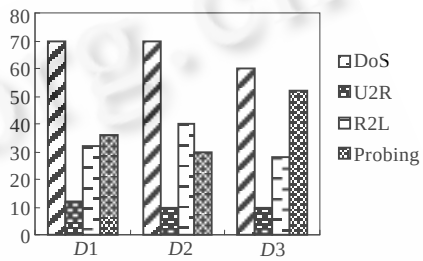

Fig.2 Attack distribution in data sets

图 2 实验数据集的攻击分布

3.2 属性选择

样本数据的 41 个属性中有 3 个标记属性和 38 个数值属性.为了对样本的属性进行评估,针对数值属性,我们训练了一个贝叶斯分类器.图 3 显示了每个属性对分类的出错率.我们把阈值设为 20%,那些出错率超过阈值的属性将被去除,结果得到 12 个属性.

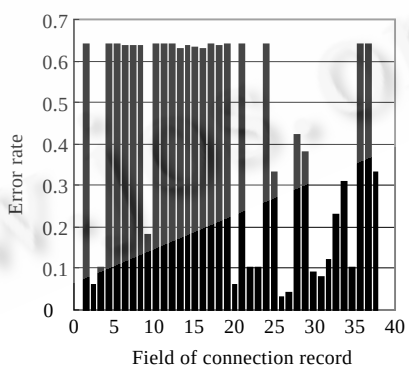

Fig.3 Error rate of numerical fields for classifying

图 3 数值属性对分类的出错率

3.3 数据预处理

对标记属性,我们进行编码处理,使其变为数值.对于数值属性,不同的属性特征有不同的度量标准,会产生大数吃小数的问题,导致数据的某些属性特征被掩盖.为了解决这个问题,必须将数据的特征属性值进行标准化,我们作如下变换[9]:

1) 计算平均的绝对偏差 S

$$S=\frac{1}{n}\sum_{i=1}^{n}(X_i-m) \tag{12}$$

其中,X_i 是数据样本,m 为样本均值,即

$$m=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{13}$$

2) 计算标准化的数据

$$Y_i=\frac{X_i-m}{S} \tag{14}$$

这相当于利用统计特性将原始实例的特征属性映射到一个标准的属性空间上,有利于减少上面所述问题.

3.4 实验结果

在实验中,选用实验平台 WindowsXP,JAVA 语言编程,Intel Pentium 1.80GHz CPU,512MB 内存.

为了验证本文的算法对入侵的检测能力,图 4 给出了 3 个数据集的平均检测率 DR(detection rate)和误检率 FR(false alarm rate)的 ROC(receiver operating characteristic)曲线图.由图中可知,最好性能当检测率为 80%时,误检率为 1.5%;而最早将聚类分析用于入侵检测的 Portnoy[10]的检测系统的检测率为 50%,误报率为 1%.这充分表明本文的算法对于未知入侵行为检测的可行性和有效性.

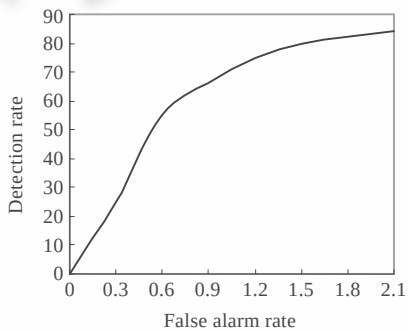

Fig.4 Detection rate vs. false alarm rate

图 4 检测率和误检率的相对曲线

为了说明本文的适应度函数在入侵检测中的有效性,我们对多种适应度函数进行了比较,结果见表 1.从比较的检测率和误警率可以看出,本文的适应度函数是有效的.

Table 1 Comparison for different fitness functions

表 1 不同适应度函数的比较

	Fitness functions	Clustering number	Detection rate (%)	False alarm rate (%)
Bezdek	$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}log_a\mu_{ij}$	12	74.0	2.68
Xie-Beni	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^2d_{ij}^2(X_i,C_j)}{n\left(\min\limits_{p\neq q}d_{pq}^2(C_p,C_q)\right)}$	35	78.0	4.3
Kwon	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^2d_{ij}^2(X_i,C_j)+\frac{1}{k}\sum_{j=1}^{k}d_j^2(C_j,\overline{C})}{\min\limits_{p\neq q}d_{pq}^2(Z_p,Z_q)}$	4	68.0	2.1
This paper	$\frac{a}{\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^md_{ij}^2(X_i,C_j)+\frac{1}{k}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}d_{ij}^2(C_i,C_j)}$	7	80.0	1.5

为了验证本文算法对确定聚类数的准确性,我们分别用本文的算法和模糊 C 均值算法对训练数据进行了实验,并用式(11)对结果进行了评价,结果如图 5 所示.从图中可以看出,模糊 C 均值算法在本文算法所确定的聚类数处使得 $f(S)$具有最大值,所以,本文的算法能够自动对数据集得出较好的聚类数(k 为本文的算法所得到的聚类数).

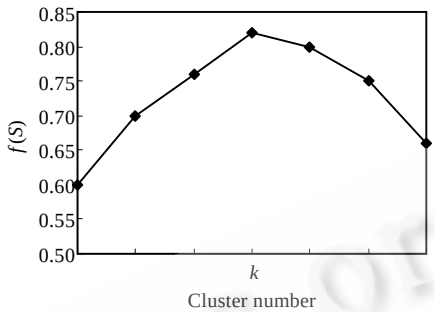

Fig.5 $f(s)$ of FCM in different clustering number

图 5 FCM 对不同的聚类数所得的 $f(s)$

针对 4 种类型的攻击,表 2 给出了各自的检测率,同时与 Lee[11]等人将数据挖掘应用于入侵检测的结果进行了比较.从表中可以看出,本文的算法在检测新型攻击方面效果良好.而对 U2R 的检测,本文的算法要差 9.9%.从图 2 可以看出,U2R 攻击总数很少,因此,这一差距实际上只有很少的攻击数,对检测结果影响不大.本文的算法对 Probing,DoS 和 U2R 都取得了较好的检测结果,但是对 R2L 的检测效果不理想,这与我们的估计基本一致.因为有很多 R2L 入侵是伪装合法用户身份进行攻击,这就使得其特征与正常数据包类似,造成了算法检测的困难.DoS 的攻击数目较多,在许多根据聚类大小进行判定的方法中需要单独处理,否则检测结果不理想,如文献[8].由于我们的判定是根据入侵数据与正常数据的差异,所以对于攻击数目较多的 DoS 攻击仍有较好的检测结果.

Table 2 Detection rate for four attacks

表 2 4 种攻击类型的检测率

Class	Detection rate for new attacks in Lee's experiments (%)	Detection rate for new attacks in our experiments (%)
Probing	96.7	91.5
DoS	24.3	88.5
U2R	81.8	71.9
R2L	5.9	52.0
Overall	37.7	80

4 结 论

在本文中,针对聚类入侵检测中聚类数难以确定的问题,提出了自动决定聚类数算法,并构建了基于该算法的入侵检测系统.通过对入侵检测数据集的实验证明,该算法实现方便,在无须监督的情况下,不仅能够得到最佳聚类数,而且对新型入侵有较好的检测效果,是一种有效的检测算法.

References:

[1] Kim DW, Lee KH. Validation of fuzzy partitions obtained through fuzzy C-means clustering. In: Zhong N, *et al*., eds. Proc. of the 14th Int'l Symp. on Foundations of Intelligent Systems. 2003. 422−426.

[2] Jiang XY, Zhao RC, Jiang ZT. Unsupervised texture segmentation based on FCM. Journal of Computer Research and Development, 2005,42(5):862−867 (in Chinese with English abstract).

[3] Li KL, Huang HK, Tian SF, Liu ZP, Liu ZQ. Fuzzy multi-class support vector machine and application in intrusion detection. Chinese Journal of Computers, 2005,28(2):274−280 (in Chinese with English abstract).

[4] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 2000. 100−101.

[5] Ben-Hur A, Horn D, Siegelmann HT, Vapnik V. Support vector clustering. Journal of Machine Learning Research, 2001,2(12): 125−137.

[6] Yang DG. Research of the network intrusion detection based on fuzzy clustering. Computer Science, 2005,32(1):86−91 (in Chinese with English abstract).

[7] Burbeck K, Nadjm-Tehrani S. ADWICE—Anomaly detection with real-time incremental clustering. In: Park C, Chee S, eds. Proc. of the 6th Int'l Conf. on Information Security and Cryptology. 2005. 407−424.

[8] Luo M, Wang LN, Zhang HG. An unsupervised clustering-based intrusion detection method. Acta Electronica Sinica, 2003,31(11): 1713−1716 (in Chinese with English abstract).

[9] Xiao LZ, Shao ZQ, Liu G. *K*-Means algorithm based on particle swarm optimization algorithm for anomaly intrusion detection. In: Proc. of the 6th World Congress on Intelligent Control and Automation. 2006. 5854−5858.

[10] Portnoy L, Eskin E, Stolfo SJ. Intrusion detection with unlabeled data using clustering. In: Proc. of the ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001). 2001. 5−8.

[11] Lee WK, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. In: Proc. of the IEEE Symp. on Security and Privacy. 1999.

附中文参考文献:

[2] 蒋晓悦,赵荣椿,江泽涛.基于 FCM 的无监督纹理分割.计算机研究与发展,2005,42(5):862−867.

[3] 李昆仑,黄厚宽,田盛丰,刘振鹏,刘志强.模糊多类支持向量机及其在入侵检测中的应用.计算机学报,2005,28(2):274−280.

[6] 杨德刚.基于模糊 C 均值聚类的网络入侵检测算法.计算机科学,2005,32(1):86−91.

[8] 罗敏,王丽娜,张焕国.基于无监督聚类的入侵检测方法.电子学报,2003,31(11):1713−1716.

肖立中(1981－),男,山东寿光人,博士,CCF 会员,主要研究领域为网络异常行为检测.

王秀英(1971－),女,博士生,讲师,主要研究领域为网络安全.

邵志清(1966－),男,博士,教授,博士生导师,主要研究领域为软件设计、开发和验证方法,网络信息服务技术

刘刚(1979－),男,博士,主要研究领域为嵌入式系统设计.

马汉华(1962－),男,博士生,高级工程师,主要研究领域为计算机网络信息安全,网络信息服务技术,计算机犯罪侦查取证.