DOI:10.7544/issn1000-1239.2015.20130601
52(3):718-728,2015

基于特征选择的模糊聚类异常入侵行为检测

唐成华[1,2] 刘鹏程[1,2] 汤申生[3] 谢 逸[4]
[1](桂林电子科技大学广西可信软件重点实验室 广西桂林 541004)
[2](桂林电子科技大学广西信息科学实验中心 广西桂林 541004)
[3](西密苏里州立大学工程技术系 美国密苏里州圣约瑟夫 64507)
[4](中山大学信息科学与技术学院 广州 510275)
(tch@guet.edu.cn)

Anomaly Intrusion Behavior Detection Based on Fuzzy Clustering and Features Selection

Tang Chenghua[1,2], Liu Pengcheng[1,2], Tang Shensheng[3], and Xie Yi[4]
[1](*Guangxi Key Laboratory of Trusted Software*, *Guilin University of Electronic Technology*, *Guilin*, *Guangxi* 541004)
[2](*Guangxi Experiment Center of Information Science*, *Guilin University of Electronic Technology*, *Guilin*, *Guangxi* 541004)
[3](*Department of Engineering Technology*, *Missouri Western State University*, *Saint Joseph*, *Missouri*, *USA MO* 64507)
[4](*School of Information Science and Technology*, *Sun Yat-sen University*, *Guangzhou* 510275)

Abstract The behaviors of network attack connection are always changeable and complex. Typical behavior mining methods, which always do using traditional clustering, do not fit in with constructing anomaly intrusion detection model. According to the characteristics of network attacks, the anomaly intrusion detection model based on fuzzy clustering and features selection are proposed. Firstly, the results that the fuzzy *C*-means clustering algorithm is sensitive to the initial cluster centers is improved using hierarchical clustering algorithm, the disadvantage that FCM is easy to fall into local optimum in the iteration is overcome using the global search ability of genetic algorithm, and they are combined into a AGFCM algorithm. Secondly, the feature attribute data sets of network attack connection are sorted through the information gain algorithm. The capacity of feature attributes is determined by using the Youden index to cut the data sets at the same time. Lastly, the anomaly intrusion detection model is built by using the attribute data sets dimensionality reduction and improved FCM clustering algorithm. Experimental results show that the anomaly intrusion detection model can effectively detect the vast majority of network attack types, which provides a feasible solution for solving the problems of false alarm rate and detection rate in anomaly intrusion detection model.

Key words fuzzy clustering; hierarchical clustering; features selection; fuzzy *C*-means; anomaly detection

摘 要 网络攻击连接具有行为的多变性和复杂性等特征,利用基于传统聚类的行为挖掘技术来构建异常入侵检测模型是不可行的.针对网络攻击行为的特点,提出了基于特征选择的模糊聚类异常入侵模型.首先通过层次聚类算法改善了FCM聚类算法结果对初始聚类中心的敏感性,再利用遗传算法的全

收稿日期:2013-04-22;**修回日期**:2014-01-13
基金项目:国家自然科学基金项目(61163057,60970146,61462020);广西可信软件重点实验室基金项目(kx201111);广西信息科学实验中心基金项目(20130329);广西自然科学基金项目(2014GXNSFAA118375)

局搜索能力克服了其在迭代时易陷入局部最优的缺点,并将它们结合构成一种 AGFCM 算法;然后采用信息增益算法对网络攻击连接数据集的特征属性进行排序,同时利用约登指数来删减数据集的特征属性以确定特征属性容量;最后利用低维特征属性集和改进的 FCM 聚类算法构建了异常入侵检测模型.实验结果表明该模型对绝大多数的网络攻击类型具有很好的检测能力,为解决异常入侵检测模型的误警率和检测率等问题提供了一种可行的解决途径.

关键词 模糊聚类;层次聚类;特征选择;模糊 C 均值;异常检测

中图法分类号 TP393

入侵检测是网络安全主动防御体系的关键技术,旨在对攻击网络和主机的行为进行检测和分类,并发出相应警报.目前常用的入侵检测技术主要分为误用检测和异常检测 2 类.早期的入侵检测技术研究集中在误用检测,通过监视目标系统的特定行为与事先定义的已知入侵模式是否匹配来判断入侵行为,具有较高的正确率、稳定的检测效果和快速响应速度,误报率低,但不能发现未知的攻击行为,需要不断更新攻击特征库,这在云计算时代仍是一个难题.异常检测是基于被保护目标的正常工作模式,将偏离正常行为的网络数据连接认定为攻击行为,能发现未知的攻击,但有较高的误报率.近年来,异常检测得到了广泛研究,在网络安全工程中发挥着越来越大的作用.

自从哥伦比亚大学的 Lee 等人[1]最早探讨了利用关联规则和序列分析等数据挖掘技术建立攻击行为挖掘模型之后,利用数据挖掘技术来挖掘数据集中的网络正常行为和攻击行为的模式开始引起国内外广泛关注.作为数据挖掘中的重要分析方法,聚类也被利用来进行入侵检测方法的研究.由于用户行为的多变性和复杂性,正常行为与异常行为的边界有时很难划定[2],对被检测数据不能简单地或硬性地归于某一类,因此往往结合多种挖掘方法分析网络行为.Denatious 等人[3]提出了可利用聚类、分类和关联规则等用于入侵行为的联合挖掘方法.Chitrakar 等人[4]组合应用 K 中心聚类算法和支持向量机算法来构建异常检测模型.Srinoy 等人[5]按照不同的隶属度综合运用粗糙模糊聚类方法实现数据容量的减少和消冗,并可用于多集群的对象,特别是对聚类间重叠区域和边界区域等进行刻画,能更客观地实现对正常行为与异常行为的软划分,实验表明这种聚类算法的结果优于已知的无监督聚类算法,由于能够描述样本类的中介性,更能客观地反映现实世界,目前已成为聚类中的主流.

作为模糊聚类中最完善的模糊 C 均值(fuzzy C-means, FCM)聚类算法,通过迭代优化目标函数得到每个样本点对所有聚类中心的隶属度,从而确定聚类数据点归属,以达到自动对样本分类的目的.但 FCM 目前尚存在诸多问题,如算法对初始聚类中心敏感、收敛速度慢、易陷入局部最优等.针对这些问题已有一些相关研究:Rose 等人[6]将模拟退火算法跳出局部最优解的能力用于聚类分析,实现全局最优,但运算时间消耗巨大,且需要满足温度下降足够慢的前提条件;张新波[7]提出改进的基于相似性阈值和最小距离原则实现粗聚类,得到各类的类中心,但在粗聚类过程中需要不断调整最小距离和相关系数,计算量大,不适合类别数目确定的聚类运算;Berget 等人[8]分析了各聚类中心间的相互影响,提出新的分离准则,更客观准确地实现聚类.还有些学者将群体智能优化算法引入聚类[9],提出基于粒子群优化的 FCM 聚类算法代替 FCM 中基于梯度下降的迭代过程,通过粒子群优化搜索得到最优初始聚类中心和适应度最优的聚类结果,但粒子群优化聚类算法对样本维数敏感,粒子群搜索空间较大,有限次迭代搜索的结果并不保证全局收敛性.

对于网络连接行为,通常利用多个特征属性加以描述,如在 KDD-Cup'99 数据集中采用 41 种特征属性来描述每条连接.但过多的维数不仅使得不相关和冗余的特征属性减慢聚类进程,而且会影响聚类的正确率,尤其是面向大数据集的特征挖掘时,通过特征选择方法找出准确刻画多结构、多类型的网络连接行为的特征,保留能够反映系统状态的重要特征对于提高入侵检测系统检测效率至关重要[10].目前已提出一些包括搜索策略和评估算法不同阶段的特征选择算法.针对特征子集空间搜索,Karimi 等人[11]提出基于多对象遗传算法的随机搜索策略,比启发式算法更有优势;陈友等人[12]提出一种基于遗传算法和禁忌搜索相混合的搜索策略进行随机搜索;Ma 等人[13]将 KDD-Cup'99 数据集特征属性分为 4 组,包括网络连接的基本特征、从主机抽取的属性特征、基于网络时间和网络流量的统计特征.实验表明选择不同的特征组用于聚类会产生不同的检测

率.在评估算法方面,Shen 等人[14]介绍了一种基于支持向量机概率灵敏度分析的特征选择算法,依据该算法可以选出那些分类信息明确的特征;Olusola 等人[15]认为在 KDD-Cup'99 入侵检测数据集中每种类型的攻击和正常行为都与数据集中各个特征属性存在不同程度的关联性,通过计算每个特征属性与各类别之间的依赖率,确定 7 种特征属性与任何攻击类型及正常类之间不存在任何关系.总之,寻找最小特征子集不仅能简化运算,且能提高速度和检测率.

基于遗传算法、聚类算法等在入侵检测模型中的应用研究已多见诸文献,如文献[16-17]单独利用聚类算法进行入侵检测;文献[18]利用聚类算法作为构建入侵检测模型过程中的数据预分类;文献[19-21]均是利用遗传算法作为优化手段以提高分类器的分类效果,各算法应用情况如表 1 所示:

Table 1 Application of Genetic Algorithm and Clustering Algorithm in Intrusion Detection Models

表 1 遗传算法和聚类算法在入侵检测模型中的应用研究情况

Reference	Algorithm	Application Description and Its Limitations	Data Set
[16]	*K*-means Clustering	The initial cluster selection, cluster number determination are accomplished by clustering center update, fission and fusion. The initial clustering distance is a new clustering inherited (offset 0) as a prerequisite to realize the new cluster label generation, to determine the different attack instances.	KDD-Cup'99
[17]	Single-linkage Clustering	No class labeled data set is clustered to construct anomaly detection model by using the single link clustering algorithm. But the calculation of the maximum cluster labels and as abnormal behavior, so as to reverse to judge the attack, brought the false alarm rate to some extent.	KDD-Cup'99
[18]	*K*-means, Naive Bayes	*K*-means is used to cluster the similar data, and then the cluster is classified using NaiveBayes, in order to build a hybrid IDS. The calculation is large because of the continuous adjustment of the correlation coefficients, and the efficiency of the algorithm is not verified.	KDD-Cup'99
[19]	GA, Fuzzy Rules	The fuzzy classification system is constructed using parallel genetic algorithm, and the fuzzy rules are generated to detect the network attacks, can solve the classification of training time consuming problem. But it default subgroups are classified using the same subjective grading fuzzy rules.	KDD-Cup'99
[20]	GA, Support Vector Machines	Genetic algorithm is used in feature selection of the experimental data, the optimal information is extracted from the raw network packets as the data preprocessing, and used to train the SVM classifier. The normal packet and its effect have not been carefully analyzed in the algorithm.	KDD-Cup'99
[21]	GA, Fuzzy Classifier	Using the genetic algorithm to construct the fuzzy classifier to describe normal and abnormal behavior characteristics can effectively reduce the false alarm rate. Pretreatment is a large amounts of calculation because of the normal and abnormal behavior detection information using the profile data set.	KDD-Cup'99

作为一直都是极具挑战性的课题,是否恰当地选取 FCM 初始聚类中心从很大程度上影响到 FCM 聚类结果,尤其在面对大数据集时随机选取初始聚类中心更会导致较高错误率. Mohammad 等人[22]提出了利用一种改进的干细胞算法(一种优化数值函数的新算法)来优化 FCM 聚类算法,并非直接选取初始聚类中心,而是通过干细胞优化算法来间接降低初始聚类中心对聚类结果所造成的影响. Ji 等人[23]提出了快速 FCM 算法进行脑 MR 图像分割,用一种自动方法来确定聚类初始中心,改善 FCM 的聚类效果以减轻灰度不均匀等不利影响.同样,张慧哲等人[24]在聚类中心选取和优化过程中充分考虑对聚类中心的约束,首先按最小距离条件初选聚类中心,再引入聚类中心之间的分离度概念,在多个域内求取聚类中心,避免 FCM 收敛结果陷入局部极小,改善了初始聚类中心对 FCM 聚类效果的产生的不利影响.

基于遗传算法的 FCM 在其他的领域中也有广泛的应用,如利用遗传算法优化 FCM 来进行图像分割. Xie 等人[25]利用基于遗传算法的 FCM 进行视网膜血管图像分割,首先得到一个近似全局最优解,再将其作为算法初始解进行 FCM 聚类以期得到全局最优解.为了克服单遗传算法在搜索最优解时的染色体数量大、进化成本高和染色体适应度的难获取等缺陷. Yoon 等人[26]提出了一种基于遗传算法的 FCM 用于解决"货郎担问题",通过计算聚簇

中代表个体的适应度,并利用 FCM 隶属度矩阵来间接求取各个染色体的适应度,降低了进化成本. Liu 等人[27]提出一种基于遗传算法的 FCM 与神经网络相结合的算法用于陀螺加速仪故障诊断,利用遗传算法较强的全局搜索能力来取得 FCM 的全局最优的聚类中心,所得最优聚类中心将作为 RBF 神经网络的输入值,实验表明该算法具有快速收敛及较高的故障诊断正确率等优点.

本文基于特征属性选择数据集,利用层次聚类算法和遗传算法来改进和优化 FCM 聚类算法,构建出的异常入侵检测模型对大多数攻击类型均有较好的检测效果. 本文首先利用基于 Agnes 层次聚类的 FCM 初始聚类中心选取算法,然后提出一种优化的 FCM 聚类算法,利用遗传操作克服在面对高维和海量数据迭代过程中易陷入局部最优及收敛速度慢的缺点,并将它们结合构成 AGFCM 算法,进而形成了一种基于特征选择和 AGFCM 的异常入侵检测模型,实现较为完善的异常入侵检测过程,并进行了实验分析和验证,最后提出了今后的工作方向.

1 FCM 聚类算法

FCM 聚类算法并非严格将每个数据点划分到某个特定类里,而是通过迭代计算使得非相似性指标的目标函数取得最小值来确定最终的聚类中心,再利用此时的隶属度矩阵来度量各个数据点相对每个类的隶属程度来确定各个数据点的归属类.

FCM 将 n 个样本数据 $X=(x_1, x_2, \cdots, x_n)$ 分为 c 个类族,并求每组的聚类中心 $v_i(i=1,2,\cdots,c)$,目标函数定义为

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m d_{ij}^2, \tag{1}$$

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

其中,u_{ij} 表示第 j 个数据点对第 i 个类的隶属度;d_{ij} 表示第 j 个数据点与第 i 个聚类中心的欧氏距离;$m\in[1,\infty]$ 是模糊聚类指数,本文取 $m=2$.

FCM 算法的过程就是最小化 J_m 的过程,在传统 FCM 算法中,需事先按经验确定聚类个数 c 和一个较小的阈值 ε. 算法每一次迭代都沿着目标函数 J_m 减少的方向进行,停止迭代的条件是相邻 2 次迭代得到的聚类中心变化很小. 由于 J_m 可能有多个极值点,而初始聚类中心的选择是随机的,因此可能使算法收敛到局部最小. 另外,J_m 只考虑每个数据点到各聚类中心的距离,不考虑各样本对聚类结果的影响和聚类中心之间的影响,导致聚类效果差.

2 FCM 聚类算法的改进

2.1 确定 FCM 聚类算法的初始聚类中心

为了克服 FCM 聚类算法对随机选取初始聚类中心敏感的缺点,利用 Agnes 层次聚类算法不需要指定初始聚类中心点,而只需初始化聚类个数这一特点来选取 FCM 聚类算法的初始聚类中心点.

本文首先对原数据集中的数据点按比例进行抽样形成原数据集的子集,记为 *SubDataSet*. 聚类个数的设定与 FCM 聚类算法保持一致. 然后利用 Agnes 聚类算法对 *SubDataSet* 数据集进行聚类得到一个与 FCM 最终聚类中心相接近的近似聚类中心 C_a,并利用 C_a 取代 FCM 聚类算法中随机选取的初始聚类中心,最后进行 FCM 聚类. 此处 Agnes 聚类算法仅是对 FCM 聚类算法在初始聚类中心的选取进行改进,并不改变 FCM 聚类算法余下的算法流程.

算法 1. Agnes 层次聚类选取初始聚类中心.

输入:原数据集的子集 *SubDataSet*、聚类中心个数 C;

输出:C 个聚类中心.

① $j=SubDataSet.Size()$; /* j 为初始簇的个数,即初始数据点的数量 */

② while ($j>C$) do

③ $C_i, C_j=FindNearestTwoCluster(Clusters)$; /* 用于寻找最近的 2 个簇,*Clusters* 为目前所有的聚类簇 */

④ $Clusters=FusionCluster(C_i, C_j)$; /* *Clusters* 为合并 2 个最近聚簇后所共有的簇集 */

⑤ $--j$;

⑥ end while

⑦ return *getClusterCenter*(*Clusters*). /* 计算各簇聚类中心以初始聚类中心返回 */

2.2 优化 FCM 聚类算法

遗传算法通过随机产生待求问题的初始解形成初始种群,每个初始解即为种群个体,再通过 3 种遗传操作来达到种群进化的目的,即通过计算种群个体适应度来选择优秀个体、种群个体间的交叉,以及种群个体本身的变异来产生新一代个体. 经过对种群个体逐代进行选择、交叉和变异操作,算法终止于收敛阈值或最大进化代数.

由于对全局最优或全局近似最优解具有较强的搜索能力,可利用遗传算法较强的全局搜索能力来克服 FCM 聚类算法在面对高维和海量数据迭代求取聚类中心点的过程中易陷入局部最优及较低搜索速度的缺点,实现对 FCM 聚类算法的快速收敛:每个种群个体独立进行 FCM 聚类算法中的适应度计算且发生在各代选择操作之前,种群的某一代经选择后得到的优秀种群个体通过交叉和变异操作产生下一代个体,即新的聚类中心点,在下一代遗传操作中通过用新的聚类中心点来更新隶属度矩阵,进而计算新种群个体的适应度,从新的聚类中心点中选择优秀个体来进行交叉和变异操作,如此迭代过程终止于最大种群进化代数.

算法 2. 基于遗传算法的 FCM 聚类算法.

输入:种群个体数 N、种群 $Population[N]$、2 个种群个体间发生交叉概率 $CrossoverProbability$、种群个体发生变异概率 $MutationProbability$、最大进化代数 $MaxGen$;

输出:全局近似最优个体 $NearOptimalMember$.

① while ($CurrentGen<MaxGen$) do
/* $CurrentGen$ 为种群当前代数 */
② for $i=0$ to ($M-1$) do
③ $UpdateMembership(Population[i])$;
/* 利用新聚类中心点更新隶属度矩阵 */
④ $Populaiton[i].fitness=CalculateFitnees(Populaiton[i])$; /* 用新隶属度矩阵计算 $1/(1+J)$,得到种群个体新的适应度; $Populaiton[N].fitness$ 为所有种群个体的适应度 */
⑤ end for
⑥ $EliteMembers=Selection(Populaiton[N].fitness)$; /* $Selection$ 为选择本代优秀种群个体 */
⑦ while ($j<EliteMembers.Size()$) do
⑧ if ($random()<CrossoverProbability$) do /* 当 $random()$ 小于交叉概率时进行交叉 */
⑨ $NewPopulation=Corssover(EliteMembers[j], EliteMembers[j+1])$; /* 产生新一代个体 */
⑩ end if
⑪ $j=j+2$;
⑫ end while
⑬ for $i=0$ to $NewPopulation.Size()-1$ do
⑭ if ($random()<MutationProbability$) do /* $random()$ 小于交叉概率时进行变异 */
⑮ $Mutation(NewPopulation)$; /* 种群个体发生变异 */
⑯ end if
⑰ end for
⑱ $Populaiton=NewPopulation$;
⑲ end while
⑳ $NearOptimalCenter=GetBestMember(Populaiton)$. /* 取得适应度最大的聚类中心 */

2.3 AGFCM 聚类算法

本文通过利用 Agnes 层次聚类算法选取 FCM 的初始聚类中心和利用遗传算法优化,此时的 FCM 聚类算法用来有针对性地克服 FCM 聚类算法的缺点,并将它们结合构成 AGFCM 算法. 在形成 AGFCM 算法后不仅克服了上文所提及的 FCM 聚类算法的缺点,还消除了在利用 Agnes 层次聚类算法进行 FCM 初始聚类中心的选取过程中,所抽取的数据点有时并不能很好地代表整个数据集的缺点,即通过利用遗传算法在初始过程中需要产生多个种群个体这一特点,保证产生能够代表原数据集的数据子集.

使用 Agnes 层次聚类算法来进行 FCM 聚类算法初始聚类中心点的选取,从全局数据集中按比例进行抽样得到全局数据集的一个子集. 通过设置聚类个数 C 进行层次聚类直到产生 C 个近似 FCM 最终聚类中心,用同样的方法进行 M 次得到初始种群,将初始种群作为基于遗传算法的 FCM 聚类算法的初始聚类中心,然后对这 M 个种群个体进行选择、交叉和变异操作.

算法 3. AGFCM 算法.

输入:种群 $Population[M]$、种群个体数 M、最大进化代数 $MaxGen$、个体间交叉概率 P_c、个体变异概率 P_m、聚类个数 C、从全局数据集中按比例抽样得到 M 个原数据集的子集 $SubDataSet$;

输出:全局近似最优个体 $NearOptimalMember$.

① for $i=0$ to ($M-1$) do
② $Population[i]=AgnesCluster(SubDataSet_i, C)$; /* 通过 $AgnesCluster$ 初始化种群个体 */
③ end for
④ $NearOptimalCenter=GFCM(Population[i], M, P_c, P_m, MaxGen)$; /* $GFCM$ 为基于

遗传算法的 FCM 聚类算法; *NearOptimalCenter* 为适应度最大的个体 */

⑤ return *NearOptimalCente*.

3 基于 AGFCM 的异常入侵检测模型

3.1 构建异常入侵检测模块

通过 AGFCM 聚类算法构建异常入侵检测模块,其构建流程如图 1 所示. 当利用 AGFCM 聚类算法对数据集进行聚类后得到正常网络连接特征的聚类中心,此聚类中心将用作异常入侵检测模块. 通过计算被检测网络连接特征与正常网络连接特征之间的聚类中心距离,并与距离阈值的大小进行比较来判定当前连接是否为网络正常连接,若大于距离阈值则将当前连接判定为网络攻击连接.

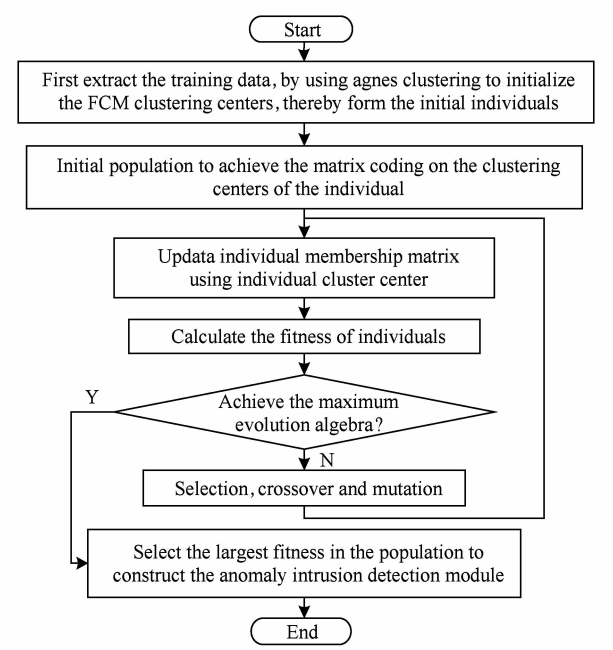


Fig. 1 Construction of anomaly intrusion detection module.

图 1 构建异常入侵检测模块的流程

3.2 异常入侵检测模型的实现

异常入侵检测模型包含以下模块:网络数据包捕获模块、协议分析模块、连接特征抽取模块、异常入侵检测模块和告警信息输出模块,如图 2 所示. 其中,网络数据包捕获模块抓取网络底层数据包. 由于入侵检测系统并不是将单个的分片数据包作为判定对象而是以一条完整的连接作为判定对象,当捕获到数据包后协议分析模块进行 IP 碎片和 TCP 流的重组,对于 IP 碎片的重组能力的大小也是衡量入侵检测系统的一项指标. 对各条完整的连接结合指定的统计时间窗、特定的网络连接数和各条连接的基本特征来进行特征抽取. 在特征抽取过程中只对信息增益排序后选取的特征进行抽取以实现降维,将这些特征信息送入异常入侵检测模块进行基于 AGFCM 的分类与检测,输出关于网络攻击连接的相关信息.

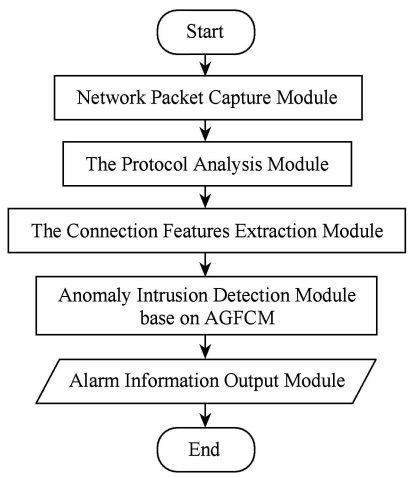


Fig. 2 Anomaly intrusion detection model.

图 2 异常入侵检测模型

4 实验过程与分析

4.1 实验数据预处理

4.1.1 数据去重处理

本文采用 KDD-Cup'99 中 10% 数据集作为模型的训练数据. 该数据集中的每条数据记录是由从一条原始网络连接中抽取的 41 维特征所组成,分别包括了 9 个基本连接特征、13 个连接内容特征和 19 个网络流量特征. 另外,采用 KDD-Cup'99 Corrected (Test)作为测试数据集,它包括 17 种在 KDD-Cup'99 中 10% 数据集所没有的攻击类型.

由于 KDD-Cup'99 的 10% 数据集有大量重复的数据记录,例如原数据集中 DOS 攻击类的 smurf 攻击记录有 280 790 条,在去除重复数据记录后仅余下 641 条. 因为算法在聚类的过程是以实例点与聚类中心点之间的距离作为计算聚簇的基础,若 smurf 的重复量过大将会在聚类过程中出现 2 种现象:1)将淹没同聚簇中其他类型的特性;2)将本应为同一簇中的实例划分到其他聚簇中. 在进行特征子集的选择时,通过除去重复网络连接记录使得在选择特征子集时能尽可能地随机抽到不同种类的连接记录,以更好地反映各数据特征对于分类所产生的影响. 去重处理后的类型分布如表 2 所示:

Table 2 The Training Data to Repeat Treatment

表 2 训练数据的去重处理

Records Number	Normal	DOS	Probe	U2R	R2L
Original Number	97 277	391 458	4 107	52	1 126
Deleted Number	9 445	336 886	1 977	0	127
Remaining Number	87 832	54 572	2 130	52	999

4.1.2 数据标准化

将数据集中的标称属性进行数值化和归一化，以消除量纲对相似度计算的影响. 归一化公式为

$$x'_{ij} = \frac{x_{ij} - x_i^{\min}}{x_i^{\max} - x_i^{\min}}, \tag{2}$$

其中，i 为数据集里第 i 个属性；j 为数据集的第 j 条记录；x_{ij} 为数据集某属性原始记录；$x_i^{\max}$ 和 $x_i^{\min}$ 分别为数据集里第 i 个属性中的最大值和最小值.

4.2 确定聚类个数

4.2.1 检测标准

本文采用计算正确率、检测率和误警率来验证入侵检测模型性能. 在对信息增益进行排序后要对特征属性进行选取来确定特征属性容量. 本文以约登指数作为选取聚类个数的依据，即通过检测率与误警率之差来确定此时聚类结果的约登指数(Youden)，作为度量分类结果优劣的标准.

4.2.2 聚类个数对聚类结果的影响

从宏观角度看，网络连接可以分为正常连接和攻击连接两类，在实际情况中攻击类网络连接又可以细分为多种，如 KDD-Cup'99 数据集中将攻击类连接分为 DOS，PROBE，U2R 和 R2L. 聚类是以相似度度量为标准进行数据点的划分，同一簇中的数据点具有最大的相似度，而不同簇中的数据点有尽可能大的相异度. 不同聚类个数下各聚簇内各网络连接类分布情况如表 3 所示. 其中当聚类个数为 2 时，C_1 和 C_2 分别表示 2 个聚簇，表 3 中数字代表 Normal，DOS，Probe，U2R 和 R2L 这 5 类连接在对应聚簇内的个数. 在 C_1 中连接种类为 Normal 的数目为 2 999，而其他 4 类连接所拥有的总个数远少于 Normal，此时 C_1 被视为 Noraml 类连接聚簇，而在 C_2 中正常类连接仅有 1 条，被视为非正常类连接聚簇.

表 3 列出了聚类个数为 2～7 时的聚类情况，从中易见一个明显的现象，即 U2R 和 R2L 一直和 Normal 类的一些连接聚在同一类中. 随着聚类个数增加，这种现象更为显著. 当聚类个数设定为 8 时(表中未列出)情况依然如此. 这是由于 U2R 和 R2L 这 2 类连接行为的特征与正常连接十分相似所导致的，这也导致在利用聚类结果对这 2 种连接的检测效果一直不明显.

Table 3 The Cluster Results in Different Clustering

表 3 不同聚类个数时的各聚簇结果

Cluster Number	Cluster	Normal	DOS	Probe	U2R	R2L
2	C_1	2 999	101	348	100	298
	C_2	1	1 899	252	0	2
3	C_1	190	367	327	0	367
	C_2	2 809	101	222	100	101
	C_3	1	1 532	51	0	1 532
4	C_1	167	366	302	0	14
	C_2	2 211	35	0	0	39
	C_3	0	1 532	52	0	0
	C_4	622	67	246	100	247
5	C_1	1 176	19	0	0	86
	C_2	435	62	247	72	72
	C_3	167	366	301	0	14
	C_4	0	1 532	52	0	0
	C_5	1 222	21	0	28	0
6	C_1	350	57	67	22	20
	C_2	206	8	180	77	210
	C_3	1 170	18	0	1	0
	C_4	163	366	296	0	14
	C_5	1 111	19	0	0	56
	C_6	0	1 532	57	0	0
7	C_1	1	1	4	0	0
	C_2	190	5	152	85	239
	C_3	360	58	91	7	1
	C_4	1 112	19	0	0	46
	C_5	160	366	295	0	14
	C_6	0	1 532	58	0	0
	C_7	1 177	19	0	8	0

在不同聚类个数的情况下得出不同的聚类中心点形成检测模型，从 KDD-Cup'99 Corrected(Test)随机抽取数据，然后依据对不同聚类个数下得到的模型进行检测以计算检测结果的正确率、检测率和误警率. AGFCM 算法参数的设定如表 4 所示：

Table 4 The Parameters of AGFCM Clustering Algorithm

表 4 AGFCM 聚类算法参数设定

Initial Population Number	Evolution Generation Number	Individuals Crossing Probability	Individuals Variation Probability	Fuzzy Index
20	40	0.8	0.1	2

将所得聚簇分别标为正常连接聚簇和异常连接聚簇,通过计算检测数据集中的网络数据连接与各个聚类中心之间的隶属度以确定该网络数据连接的属类.对于各个聚簇,当簇中正常类连接数目大于攻击类时就设定该簇为正常类聚簇.

通过对表 5 和表 6 分析可知,随着聚类个数的增加,对攻击类的检测能力逐渐增强,但是对正常类检测的正确率却在下降即误警率在增加.通过计算约登指数并对比实验结果,本文的聚类个数设置为 2,但此时对于 U2R 和 R2L 的检测率却很低,可在日常生活中以正常连接为多数,若是增加聚类个数虽然提高了 U2R 和 R2L 的检测率,却会导致误警率的大幅提高.

Table 5 The Connection Test Results

表 5 不同聚类个数时的各类连接检测结果 %

Cluster	Correct Rate				
	Normal	DOS	Probe	U2R	R2L
C_2	99.9	94.8	42.7	0	0.5
C_3	93.7	94.9	62.9	0	4.6
C_4	74.6	95.2	99.9	100	84.9
C_5	81.4	98.3	99.9	64	30.9
C_6	88	94.9	90	74	75.5
C_7	87.5	95	89.4	79.4	79.4

Table 6 The Test Results of Different Clustering

表 6 不同聚类个数下的检测结果

Cluster	Correct Rate/%	Detection Rate/%	False Alarm Rate/%	Youden Index
C_2	85.8	99.9	0.02	0.998
C_3	85	92.4	0.06	0.923
C_4	86	79.3	25.4	0.539
C_5	86	82.9	18.6	0.644
C_6	89.6	88	11.3	0.767
C_7	89.6	88	12.4	0.756

4.3 特征子集的选择

通过上述实验确定了聚类的个数,进而对数据集进行特征属性选择以去除冗余分类特征属性保留分类能力强的特征属性,降低数据特征容量从而降低运算复杂度提高效率.本文利用信息增益排序进行特征属性选择,以确定最后的属性空间.不同数据特征具有不同的信息增益,具有较大信息增益的特征属性有更强的分类能力.通过计算信息增益量得出特征属性对分类的重要性,从全局衡量特征属性对分类的贡献.对网络攻击连接的特征属性的信息增益计算方法如下.

1) 计算特征数据集 D 的熵:

$$H(D)=-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\times \mathrm{lb}\,\frac{|C_k|}{|D|}\right). \tag{3}$$

2) 计算特征属性 A 对数据集 D 的条件熵:

$$H(D\mid A)=-\sum_{i=1}^{n}\left(\frac{|C_i|}{|D|}\times\sum_{k=1}^{K}\left(\frac{|C_{ik}|}{|D_i|}\times \mathrm{lb}\,\frac{|C_{ik}|}{|D_i|}\right)\right). \tag{4}$$

3) 计算信息增益:

$$InfoGain(D,A)=H(A)-H(D\mid A). \tag{5}$$

式(3)~(5)中,D 为数据集;A 为特征属性;$|D|$为数据集样本容量;k 为所有类种数;$|C_k|$表示属于 k 类的样本容量,依据特征属性 A 的取值将 A 划分为 n 个子集 $D_1,D_2,\cdots,D_n$,将 D_i 样本容量记为$|D_i|$,子集 D_i 中类 C_k 的样本容量记为$|D_{ik}|$.

利用特征属性信息增益的计算方法对聚类所用的数据集进行特征选取:从 KDD-Cup'99 10%数据集中随机抽取去重后的数据 1 000 条正常网络连接和 1 000 条攻击网络连接,其中这 1 000 条攻击网络连接通过分层抽取 4 类攻击中的 22 种攻击类型.通过此抽取方法构建 5 组数据集,然后通过对各数据集进行信息增益计算及排序,取得每种特征属性信息增益的平均值并进行降序排序.表 7 为构建的数据子集中各个特征属性的平均信息增益值的降序排序:

Table 7 The Training Data Set Attribute Information Gain Value

表 7 训练数据集特征属性的信息增益值

Attribute Name	Gain Value	Youden Index	Attribute Name	Gain Value	Youden Index
src_bytes	2.312 569	0	srv_diff_host_rate	0.409 955	0.995 8
dst_bytes	1.878 287	0	same_srv_rate	0.376 468	0.999 2
service	1.800 958	0	num_file_creations	0.366 738	0.999 2
dst_host_srv_count	1.290 563	0.619	diff_srv_rate	0.362 988	0.979 8
dst_host_same_src_port_rate	0.923 855	0.619	srv_rerror_rate	0.355 727	0.979 8

Table 7 (Continued)

Attribute Name	Gain Value	Youden Index	Attribute Name	Gain Value	Youden Index
duration	0.921 755	0.565	rerror_rate	0.346 644	0.974 2
dst_host_same_srv_rate	0.917 242	0.571 2	srv_serror_rate	0.337 947	0.974 5
dst_host_diff_srv_rate	0.851 552	0.571 2	root_shell	0.329 253	0.974 6
count	0.848 228	0.633 9	num_compromised	0.308 558	1
dst_host_count	0.805 776	0.637 2	wrong_fragment	0.305 123	1
srv_count	0.798 654	0.66	num_root	0.291 279	1
dst_host_serror_rate	0.795 853	0.698	num_shells	0.231 451	1
dst_host_srv_serror_rate	0.706 231	0.698	num_access_files	0.222 804	1
flag	0.681 29	0.789 8	land	0.153 084	1
dst_host_srv_diff_host_rate	0.675 689	0.864 5	num_failed_logins	0.145 517	1
dst_host_rerror_rate	0.620 537	0.988 5	su_attempted	0.072274	1
hot	0.589 072	0.987 7	is_guest_login	0.068 475	1
dst_host_srv_rerror_rate	0.491 415	0.993	urgent	0.040 746	1
protocol_type	0.483 591	0.99	num_outbound_cmds	0.0	1
serror_rate	0.468 656	0.993 8	is_host_login	0.0	1
logged_in	0.444 156	0.993 8			

当依次删除全部特征属性后比对不同特征属性容量下约登指数的大小,进行全局度量选择约登指数下降不大且具有较少特征子空间的那组结果所对应的特征属性容量.实验最后选取的特征属性为前22个特征属性.(表7中约登指数是指当删除当前属性后模型检测结果的约登指数).

实验所用数据集共分为4组,每组分别包含10 000,20 000,30 000,40 000条数据记录.利用新特征子集运行AGFCM算法与未筛选特征子集时的运行时间进行比较,如图3所示:

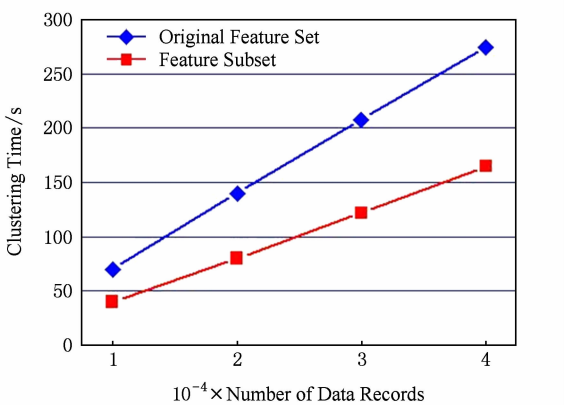


Fig. 3 Time comparison for clustering.

图3 筛选特征子集前后聚类所用时间对比

4.4 算法性能对比

本文利用KDD-Cup'99中10%数据集进行入侵检测模型的训练,选择数据集中的特征子集进行基于AGFCM聚类,并在KDD-Cup'99 Corrected(Test)中随机抽样形成的5组测试集中进行检测.按照表4设置参数,表8则展示了在同样的训练数据集和5组测试集的情况下,FCM,NaiveBayse和AGFCM算法的检测结果对比:

Table 8 Detection Results of Different Algorithms

表8 算法检测结果对比 %

Test Set	Comparison Parameter	FCM	NaiveBayse	AGFCM
Test Set 1	Detection Rate	93.23	97.15	99.85
	False Alarm Rate	7.0	18.3	0.013
Test Set 2	Detection Rate	93.47	97.23	99.85
	False Alarm Rate	6.8	17.77	0.013
Test Set 3	Detection Rate	93.6	97.33	99.89
	False Alarm Rate	6.7	18.06	0.01
Test Set 4	Detection Rate	94.7	97.48	99.92
	False Alarm Rate	5.7	16.9	1.5
Test Set 5	Detection Rate	93.68	97.1	99.89
	False Alarm Rate	6.6	18.1	0.1
The Average Detection Rate		93.71	97.26	99.89
The Average False Alarm Rate		6.6	17.83	0.010 6

可以看出,AGFCM较FCM和NavieBayes在入侵检测应用中有更高的检测率和更低的误警率.

5 结 论

FCM是通过隶属度矩阵对数据集中网络连接数据进行软划分,本文有针对性的将层次聚类引入到FCM聚类算法的初始聚类点的选取中,通过遗传算法提高了FCM算法对聚类中心的全局搜索的能力,利用信息增益算法实现数据集特征属性的排序,同时结合不同特征属性容量下的约登指数值来删减冗余分类属性提高了聚类算法速度.实验表明在降低特征集容量的情况下,仍能达到较好的聚类效果,为解决特征子集的选取及利用FCM取得更好的聚类效果等问题提供了一种可行的思路.实验表明本文提出的算法是可行的和有效的,对于绝大多数网络攻击连接具有很好的检测效果.

参 考 文 献

[1] Lee W, Stolfo S J. Data mining approaches for intrusion detection [C] //Proc of the 7th USENIX Security Symposium. Berkeley, CA: USENIX Association, 1998: 79-93

[2] Li Chao, Tian Xinguang, Xiao Xi, et al. Anomaly detection of user behavior based on shell commands and co-occurrence matrix [J]. Journal of Computer Research and Development, 2012, 49(9): 1982-1990 (in Chinese)
(李超, 田新广, 肖喜, 等. 基于Shell命令的共生矩阵的用户行为异常检测方法[J]. 计算机研究与发展, 2012, 49(9): 1982-1990)

[3] Denatious D K, John A. Survey on data mining techniques to enhance intrusion detection [C] //Proc of the 2012 Int Conf on Computer Communication and Informatics. Piscataway, NJ: IEEE, 2012: 1-5

[4] Chitrakar R, Huang Chuanhe. Anomaly detection using support vector machine classification with *K*-medoids clustering [C] //Proc of the 3rd Asian Himalayas Int Conf on Internet. Piscataway, NJ: IEEE, 2012:1-5

[5] Srinoy S, Kurutach W, Chimphlee W, et al. Intrusion detection via independent component analysis based on rough fuzzy [J]. WSEAS Trans on Computers, 2006, 5(1): 43-48

[6] Rose K, Gurewitz E, Fox G C. Constrained clustering as optimization method [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1993, 15(8): 785-794

[7] Zhang Xinbo. Fuzzy *C*-means clustering algorithm with two stages [J]. Journal of Circuits and System, 2005, 10(2): 117-121 (in Chinese)
(张新波. 两阶段模糊*C*-均值聚类算法[J]. 电路与系统学报, 2005, 10(2): 117-121)

[8] Berget I, Mevik B H. New modifications and applications of fuzzy *C*-means methodology [J]. Computational Statistics and Data Analysis, 2008, 52(5): 2403-2418

[9] Wang Zonghu, Liu Zhijing, Chen Donghui. Research of PSO-based fuzzy *C*-means clustering algorithm [J]. Computer Science, 2012, 39(9): 166-169 (in Chinese)
(王纵虎, 刘志镜, 陈东辉. 基于粒子群优化的模糊*C*-均值聚类算法研究[J]. 计算机科学, 2012, 39(9): 166-169)

[10] Chen You, Cheng Xueqi, Li Yang, et al. Lightweight intrusion detection system based on feature selection [J]. Journal of Software, 2007, 18(7): 1639-1651 (in Chinese)
(陈友, 程学旗, 李洋, 等. 基于特征选择的轻量级入侵检测系统[J]. 软件学报, 2007, 18(7): 1639-1651

[11] Karimi N M, Konstantaras I. A random search heuristic for a multi-objective production planning [J]. Computers & Industrial Engineering, 2012, 62(2): 479-490

[12] Chen You, Shen Huawei, Li Yang, et al. An efficient feature selection algorithm toward building lightweight intrusion detection system [J]. Chinese Journal of Computers, 2007, 30(8): 1398-1408 (in Chinese)
(陈友, 沈华伟, 李洋, 等. 一种高效的面向入侵检测系统的特征选择算法[J]. 计算机学报, 2007, 30(8): 1398-1408)

[13] Ma Wanli, Tran D, Sharma D. A study on the feature selection of network traffic for intrusion detection purpose [C] //Proc of IEEE ISI'08. Piscataway, NJ: IEEE, 2008: 245-247

[14] Shen Kaiquan, Ong Chongjin, Li, Xiaoping, et al. Feature selection via sensitivity analysis of SVM probabilistic outputs [J]. Machine Learning, 2008, 70(1): 1-20

[15] Olusola A A, Oladele A S, Abosede D O. Analysis of KDD-Cup'99 intrusion detection dataset for selection of relevance features [C] //Proc of the 2010 World Congress on Engineering and Computer Science. Piscataway, NJ: IEEE, 2010: 162-168

[16] Jungsuk S, Kenji O, Hiroki T. A clustering method for improving performance of anomaly-based intrusion detection system [J]. IEICE Trans on Informatin and Systems, 2008, E91-D(5): 1282-1291

[17] Portony L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering [C] //Proc of ACM DMSA-2001. New York: ACM, 2001: 123-130

[18] Muda Z, Yassin W, Sulaiman M N, et al. A *K*-means and Naïve Bayes learing approach for better intrusion detection [J]. Information Technology Journal, 2011, 10(3): 648-655

[19] Abadeh M S, Habibi J, Barzegar Z, et al. A parallel genetic local search algorithm for intrusion detection in computer networks [J]. Engineering Applications of Artificial Intelligence, 2007, 20(8): 1058-1069

[20] Taeshik S, Jongsub M. A hybrid machine learning approach to network anomaly detection [J]. Information Sciences, 2007, 177(18):3799-3821

[21] Gomez J, Dasgupta D, Nasraoui O, et al. Complete expression trees for evolving fuzzy classifier systems with genetic algorithms [C] //Proc of IEEE NAFIPS'02. Piscataway, NJ: IEEE, 2002: 469-474

[22] Mohammad T, Mohammad H B. A powerful hybrid clustering method based on modified stem cells and fuzzy C-means algorithms [J]. Artificial Intelligence, 2013, 26(5/6): 1493-1502

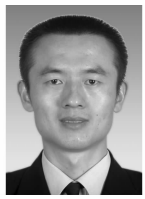
[23] Ji Zexuan, Sun Quansen, Xia Deshen. A framework with modified fast FCM for brain MR images segmentation [J]. Pattern Recognition, 2011, 5(44): 999-1013

[24] Zhang Huizhe, Wang Jian. Improved fuzzy C-means clustering algorithm based on selecting initial clustering centers [J]. Computer Science, 2009, 36(6): 206-209 (in Chinese)
(张慧哲,王坚. 基于初始聚类中心选取的改进 FCM 聚类算法[J]. 计算机科学, 2009, 36(6): 206-209)

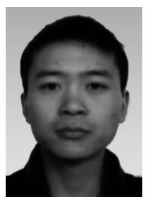
[25] Xie Songhua, Nie Hui. Retinal vascular image segmentation using genetic algorithm plus FCM clustering [C] //Proc of the 3rd Int Conf on Intelligent System Design and Engineering Applications. Piscataway, NJ: IEEE, 2013: 1225-1228

[26] Yoon J W, Cho S B. An efficient genetic algorithm with fuzzy C-means clustering for traveling salesman problem [C] //Proc of the 2011 IEEE Congress on Evolutionary Computation. Piscataway, NJ: IEEE, 2011: 1452-1456

[27] Liu Zhide, Chen Jiabin, Song Chunlei. A new RBF neural network with GA-based fuzzy C-means clustering algorithm for SINS fault diagnosis [C] //Proc of IEEE CCDC'09. Piscataway, NJ: IEEE, 2009: 208-211

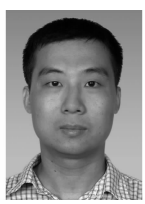


Tang Chenghua, born in 1974. Received his PhD degree in 2007 from the Beijing Institute of Technology. His main research interests include information security, data mining, and security policy analysis.



Liu Pengcheng, born in 1987. Master candidate. His current research interests include network information security, data mining, etc.

Tang Shensheng, born in 1969. PhD. His research interests include communication systems and networking, network security behavior analysis, intelligent information processing.



Xie Yi, born in 1973. PhD. His research interests include detection and analysis of network behavior, data mining.