基于人工鱼群的优化K-means聚类算法

于海涛[1,2] 贾美娟[1,2] 王慧强[2] 邵国强[1]
(大庆师范学院计算机科学与信息技术学院 大庆 163712)[1]
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)[2]

摘 要 针对K-means算法全局搜索能力不足,提出基于人工鱼群的优化K-means聚类算法(AFS-KM),该算法克服了K-means聚类算法对初始聚类中心选择的敏感问题,能够获得全局最优的聚类划分。在聚类过程中,采用一种基于信息增益的属性加权的实体之间距离计算方法进行聚类划分时,对于球形数据和椭球形数据都能够获得理想的聚类划分结果。对KDD-99数据集的仿真实验结果表明,该算法在网络入侵检测时获得了理想的检测率和误报率。

关键词 聚类,人工鱼群,信息增益,属性加权,入侵检测
中图法分类号 TP3-05 **文献标识码** A

K-means Clustering Algorithm Based on Artificial Fish Swarm

YU Hai-tao[1,2] JIA Mei-juan[1,2] WANG Hui-qiang[2] SHAO Guo-qiang[1]
(College of Computer Science and Information Technology, Daqing Normal University, Daqing 163712, China)[1]
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[2]

Abstract Aimed at the lack of global search capability of K-means algorithm, optimized K-means clustering algorithm based on artificial fish swarm(AFS-KM) was presented in this paper, which can overcome the problem of initial clustering center selection sensitivity of K-means and can obtain global optimized clustering partition. During clutering process, a weighted distance computation method based on information gain attribute weighting is used, so, the better clustering can be obtained for both spherical data and ellipsodal data. Simulation experiment is implemented over data set KDD-99, and the result shows that the satisfying detection rate and false acceptance rate can be obtained in network intrusion detection.

Keywords Clustering, Artificial fish swarm, Information gain, Attribute weighting, Intrusion detection

聚类分析[1]是数据挖掘中的一个重要研究方向,它对数据不做任何统计假设,在模式识别和人工智能等领域称为无监督学习。目前聚类分析广泛应用于网络入侵检测、医学图像处理、文本检索、生物信息学等研究领域。经典的聚类算法有K-means算法、K-原型算法、模糊聚类算法。K-means算法针对具有连续特征属性的数值型数据进行聚类划分,具有较好的伸缩性,从而被广泛采用;K-原型算法在进行聚类时增加了离散型分类特征属性的数值化,对于离散属性之间的距离采用简单的0-1匹配方法进行计算,该方法简单但是弱化了类内的相似性;为此文献[2]提出基于粗糙集的K-modes算法,该方法充分体现了分类属性在计算距离时所起的作用,但是粗糙集方法计算属性权值的精度较低;模糊聚类算法在聚类的过程中将样本划到多个聚类中而不是一个聚类中,广泛应用于模糊控制领域。但是这些聚类算法都是随机选择初始聚类中心,对初始的聚类中心没有进行任何优化,因此不容易获得最优的聚类划分。若初始中心选择不合适,聚类算法就容易陷入到局部最优解,不能获得全局近似最优的聚类划分,并且算法的收敛时间较长,应用这些聚类算法进行分类时,不容易获得令人满意的分类效果。为了解决聚类初始中心选择敏感问题,文献[3]将模拟退火算法和聚类算法相结合对聚类算法进行优化,但是模拟退火算法需要调整许多参数,人为因素可能造成聚类结果的差异,需要通过大量的数值模拟才能选择比较好的参数搭配,从而影响了算法的推广。为此,文献[4,5]提出全局优化的方法来进行聚类优化,但是这些方法仅仅考虑了进化操作的创新,在计算欧式几何距离时,隐含假定聚类的训练数据和测试数据的各维特征属性对于聚类的贡献相同,没有考虑各个特征属性对于聚类的影响,各特征属性的贡献相同不能真实反映数据的特征分布,对于椭球形数据不能获得理想的分类效果。而在实际应用中,构成训练数据和测试数据的各维特征属性来自不同的传感器,存在量纲和精度的差异,所以以上聚类算法在实际应用中具有一定的局限性。为了提高聚类的分类效果,文献[6]利用核方法将聚类初始数据映射到高维的稀疏数据空间,以提高数据的线性可分,此方法虽然能够提高数据的聚类效果,但是对于高维数据容易造成维数灾难。文献[7]将流行算法中的流行距离测度方法和遗传算法结合起来对聚类进行优化,流行距离

到稿日期:2012-06-04 返修日期:2012-09-24 本文受黑龙江省教育厅科技研究项目(11553001)资助。
于海涛(1973-),男,博士生,讲师,主要研究方向为网络安全,E-mail:albertyht@163.com;贾美娟(1976-),女,博士生,讲师,主要研究方向为网络与信息安全;王慧强(1960-),男,教授,博士生导师;邵国强(1981-),男,硕士,讲师,主要研究方向为网络安全。

计算方法对于分布复杂的非欧氏几何空间中的数据效果较好，但是对于欧式几何空间中的数据，采用流行距离计算方法，其时间复杂度较高，所以通用性不强。

针对以上问题，本文提出了基于人工鱼群的优化 K-means 聚类算法，人工鱼群算法具有先天的并行性，而且对于参数选择具有较好的健壮性，解决了聚类对初始中心选择敏感的问题，能够获得全局最优的聚类划分，从而在实际应用中能获得理想的分类效果。在聚类的过程中，采用信息增益的加权距离计算方法，充分体现了各个属性在计算距离时所起的作用，属性加权的距离计算方法对于球形数据和椭球形数据都能获得理想的分类效果。本文第 1 节介绍基于信息增益的加权距离计算方法；第 2 节详细描述了基于人工鱼群的优化 K-means 聚类算法；第 3 节是仿真实验结果；最后是全文的结论。

1 基于信息增益的加权距离计算方法

信息增益[8]方法是根据特征属性在分类过程中携带的信息量来确定该属性的分类贡献率，某个特征属性所携带的信息量越大，那么该特征属性在分类中起到的贡献就越大，则该属性的权值越大，反之亦然。

对于网络入侵检测问题，设数据集 $X=\{x_1,x_2,\cdots,x_N\}\in R^q$，$x_i=(x_{i,1},x_{i,2},\cdots,x_{i,q})$，属性集合为 $A=\{A_1,A_2,\cdots,A_q\}$，$A_i\in\{A_{i,1},A_{i,2},\cdots,A_{i,k}\}$，其中 $x_{i,k}$ 表示第 i 个实体的第 k 个特征属性，在训练集合中满足 $y_i=f(x_i)$，其中 $y_i\in\{C_1,C_2,\cdots,C_s\}$，$y_i$ 是样本 x_i 所属的类别。

1)分类系统的熵的计算公式为：

$$H(C)=-\sum_{i=1}^{s}P(C_i)\times\log_2 P(C_i) \tag{1}$$

2)特征 A_i 的信息增益计算公式为：

$$IG(A_i)=H(C)-H(C|A_i) \tag{2}$$

3)$H(C|A_i)$为属性 A_i 时的条件熵，计算公式为：

$$H(C|A_i)=P(A_i)\times H(C|A_i)+P(\overline{A}_i)\times H(C|\overline{A}_i) \tag{3}$$

式中，$P(A_i)$为特征 A_i 出现时的概率，$P(\overline{A}_i)$为特征属性不出现时的概率，$H(C|A_i)$为特征属性 A_i 出现时的条件熵，$H(C|\overline{A}_i)$为特征属性 A_i 不出现时的条件熵，计算公式如下：

$$H(C|A_i)=-\sum_{j=1}^{s}P(C_j|A_i)\times\log_2 P(C_j|A_i),H(C|\overline{A}_i)$$
$$=-\sum_{j=1}^{s}P(C_j|\overline{A}_i)\times\log_2 P(C_j|\overline{A}_i)$$

式中，$P(C_j|A_i)$为特征属性 A_i 出现时类别 C_j 出现的概率；$P(C_j|\overline{A}_i)$为特征属性 A_i 不出现时类别 C_j 出现的概率。

4)第 k 个特征属性的贡献率为：

$$\gamma_k=IG(A_i)/(\sum_{i=1}^{q}IG(A_i)) \tag{4}$$

γ_k 为第 k 个特征属性的贡献率，贡献率表示该属性在聚类或分类过程中所起的贡献的大小，即在计算实体之间距离的属性权值。获得特征属性的权值，实体 i 和实体 j 之间的距离为 $Dist(x_i,x_j)=\sum_{k=1}^{q}\gamma_k^2\cdot(x_{i,k}-x_{j,k})^2$。这里没有采用简单的欧式几何距离公式，而是采用了一种改进的欧式几何距离公式，该距离公式没有采用开方运算，直接使用平方和的累加运算，这样可以显著地减少计算量，而不影响实体之间距离的度量，提高了算法的运算速度，降低了算法的计算复杂性。

2 基于人工鱼群的优化 K-means 聚类算法

人工鱼群算法[9]是一种基于鱼类觅食行为的智能优化算法，采用自上而下的设计方法，对寻优空间的形式没有特殊要求，具有较好的全局寻优能力。人工鱼群算法在寻优的过程中，因为采用群体合作方式和概率方法进行状态转移，所以具有先天的并行性，并且能够避免在局部最优解的振荡迂回搜索，从而极大地减少了算法的收敛时间。人工鱼群算法对初值、参数选择具有较好的健壮性，具有良好地克服局部最优解并获得全局最优解的能力，能够找到满意解的范围，因此适用于寻找 K-means 算法初始聚类中心。但人工鱼群算法后期的收敛速度较慢，很难得到精确的最优解，而 K-means 在获得较好的初始聚类中心后，经过较少的迭代次数就能获得最优的聚类划分。基于人工鱼群的优化 K-means 聚类算法结合了人工鱼群算法和 K-means 算法的优点，既能克服 K-means 算法对于初始聚类中心选择敏感的问题，又能解决人工鱼群后期收敛速度较慢的问题，因而能够在相对较短的时间内获得最优的聚类划分。

2.1 聚类的有关定义

为便于说明本文算法，对聚类算法作如下具体定义。设本文采用的数据为 $X=\{x_1,x_2,\cdots,x_N\}$，$x_j\in R^q$。

定义 1(聚类的有关参数) k 为聚类个数；聚类划分结果 $C=\{class_i|\ i\leqslant k\}$，$class_i$ 为第 i 个聚类集合；$S=\{center_i|\ i\leqslant k\}$，$center_i$ 为 $class_i$ 的聚类中心。

定义 2 类内间距表示一个类的数据分布的紧密程度。类内间距定义为：

$$dist\text{-}within\text{-}class=\sum_{i=1}^{k}\sum_{x_l\in class_i}\sqrt{\sum_{j=1}^{q}(x_{l,j}-\overline{x_{i,j}})^2}$$

式中，$\overline{x_{i,j}}=\frac{1}{|class_i|}\sum_{x_l\in class_i}x_{l,j}$，$|class_i|$为集合中元素的个数。

定义 3 类间间距表示划分聚类的不同类别之间的距离。类间距离定义为：

$$dist\text{-}among\text{-}class=\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}|center_i-center_j|$$

式中，$|center_i-center_j|$为聚类中心的欧式距离。

定义 4 聚类的评价函数表示为聚类划分结果的优劣。聚类 C 的评价函数定义为：

$$Fitness(C)=\frac{dist\text{-}among\text{-}class}{dist\text{-}within\text{-}class}$$

2.2 人工鱼的有关定义

为了利用人工鱼算法优化聚类算法，需要对人工鱼算法中的有关概念进行具体定义。

定义 5(人工鱼的有关参数) 人工鱼的定义为 $Fish^i=(Fish_1^i,Fish_2^i,\cdots,Fish_k^i)$，表示第 i 条人工鱼所处的空间状态，其中 $Fish_j^i=(Fish_{j,1}^i,Fish_{j,2}^i,\cdots,Fish_{j,q}^i)$为第 i 条人工鱼的第 j 个聚类中心，此时人工鱼处于 $q*k$ 维空间中；人工鱼视野范围用 V 表示；人工鱼移动的最大步长为 STEP，try-number 表示人工鱼每次觅食最大的尝试次数。

定义 6 人工鱼之间的距离表示在寻优过程中两条人工鱼在当前所处状态空间中的欧式距离。人工鱼 $Fish^i$ 与 $Fish^j$ 之间的距离为 $D(Fish^i,Fish^j)=\sqrt{\sum_{s=1}^{k}\sum_{t=1}^{t=q}(Fish_{s,t}^i-Fish_{s,t}^j)^2}$。

定义 7 食物浓度为人工鱼当前所处位置的食物浓度，

在本文算法中采用聚类的划分结果表示人工鱼所在位置的食物浓度。第 i 条鱼所处位置的食物浓度定义为 $food(Fish^i)=Fitness(C^i)$，其中 C^i 表示第 i 条人工鱼当前所处状态空间中的聚类划分结果。

定义 8 K 近邻表示距离小于 K 的人工鱼。对于人工鱼 $Fish^i$，如果 $D(Fish^i,Fish^j)\leqslant K$，则 $Fish^j$ 称为 $Fish^i$ 的一个 K 近邻。

定义 9 聚群鱼群的中心位置表示聚群人工鱼所在状态空间中的欧式几何的中心位置。设 G 为聚群人工鱼的集合，$G=\{Fish^1,Fish^2,\cdots,Fish^m\}$，人工鱼 $Fish^1,Fish^2,\cdots,Fish^m$ 的中心位置(聚群的鱼群的中心位置)定义为 $Center(Fish^1,Fish^2,\cdots,Fish^m)$，如果数值类型的状态分量为 $Mean(Fish^1,Fish^2,\cdots,Fish^m)$，$Mean$ 函数表示 m 条人工鱼的相应分量的平均值；如果分类类型的人工鱼状态为 $Most(Fish^1,Fish^2,\cdots,Fish^m)$，$Most$ 函数表示 m 条人工鱼中大多数人工鱼所取的分量值操作。

2.3 基于人工鱼群的优化聚类算法的实现步骤

基于人工鱼群的优化聚类算法首先初始化人工鱼群，并将当前的最优的食物浓度记录在公告板上；然后利用模拟人工鱼群的觅食行为、聚群行为、追尾行为和随机行为并行地智能搜索聚类的初始中心；最后在利用人工鱼群的算法获得的优化聚类中心运行 K-means 算法。

算法 1 基于人工鱼群的优化聚类算法

输入：$X=\{x_1,x_2,\cdots,x_N\}$，$x_j\in R^q$，聚类个数 k

输出：X 的 k 个最优的聚类划分 C，k 个聚类的中心 $S=\{center_1,center_2,\cdots,center_k\}$

Step 1 初始化 n 条人工鱼：

对于每条人工鱼执行如下操作：

1.1 $Fish^i=\{center_1,center_2,\cdots,center_k\}$ = Random_select(X, k)// 随机选择 k 个实体作为初始的聚类中心，初始聚类中心作为 S_i 第 i 条人工鱼所在的位置

1.2 C_i = Partition($Fish^i$)//X 中实体按照最短距离原则进行聚类划分

1.3 $Fish^i$ = Center(C_i)；更新人工鱼的位置(聚类中心)

1.4 $food_i$ = food($Fish^i$)；

1.5 $food_{best}$ = max($food_i$)；$Fish_{best}$ = $Fish'$；//最大的食物浓度及其所在的位置记录在公告板上

Step 2 利用人工鱼群算法搜索初始的聚类中心

While($food_{best}<\varepsilon$) 对于第 i 条人工鱼($1\leqslant i\leqslant n$)执行如下步骤：

2.1 觅食行为：

While(number<trynumer)

{ For(j=1;j<=k;j++)

For(L=1;L<=q;L++)

{$\Delta Fish^j_{j,L}$ = Random(−1,+1) * step; $Fish'^j_{j,L}=Fish^j_{j,L}+\Delta Fish^j_{j,L}$;}

If(food($Fish^i$)<food($Fish'$) && D($Fish^i$,$Fish'$)<V)

{$Fish^i=Fish'$;

C_i = Partition($Fish^i$); $Fish^i$ = Center(C_i);

if($food_{best}$< food($Fish^i$)) {$food_{best}$ = food($Fish^i$); $Fish_{best}$ = $Fish^i$;}

break;}

Else number++;

}

2.2 聚群行为：

set=Φ;

For(j=1;j≤n;j++)

If(D($Fish^i$,$Fish^j$)<V && food($Fish^i$)<food($Fish^j$)) set=set∪$Fish^j$

$Fish^i=\frac{1}{|set|}\sum_{Fish^j\in set}Fish^j$; C_i = Partition($Fish^i$); $Fish^i$ = Center(C_i);

if(food($Fish^i$)>$food_{best}$)

{$food_{best}$ = food($Fish^i$); $Fish_{best}$ = $Fish^i$;}

2.3 追尾行为：

Max-food=food($Fish^i$);

$Fish'=Fish^i$;

For(j=1;j≤n;j++)

If(D($Fish^i$,$Fish^j$)<V && max-food<food($Fish^j$))

{max-food=food($Fish^j$); $Fish'=Fish^j$;}

If($Fish'$! =$Fish^i$)

{$Fish^i=\frac{Fish'-Fish^i}{}\times step$; C_i = Partition($Fish^i$); $Fish^i$ = Center(C_i);

if(food($Fish^i$)>$food_{best}$)

{$food_{best}$ = food($Fish^i$); $Fish_{best}$ = $Fish^i$;}

}

2.4 随机行为：

For(j=1;j<=k;j++)

For(L=1;L≤q;L++)

{$\Delta Fish^j_{j,L}$ = Random(−1,+1) * step; $Fish^j_{j,L}=Fish^j_{j,L}+\Delta Fish^j_{j,L}$;}

C_i = Partition($Fish^i$); $Fish^i$ = Center(C_i);

if(food($Fish^i$)>$food_{best}$) {$food_{best}$ = food($Fish^i$); $Fish_{best}$ = $Fish^i$;}

Step 3

3.1 C=K-means(X, $Fish_{best}$, k);// 使用各聚类的初始中心集合，运行 K-means 算法，C 为最优聚类划分；

3.2 S=Center(C);//获得最终聚类划分更新聚类中心集合

参数说明：Partition($Fish^i$)根据人工鱼当前的位置，按照最短距离原则进行聚类划分；Center(C_i)根据聚类划分结果更新聚类中心，然后用最新的聚类中心更新人工鱼所在的位置；Random(−1,+1)函数取值[−1,+1]之间的随机数，这样可以保证人工鱼既能前进也能后退；本算法中人工鱼在到达新的位置后，按照最短距离原则对数据集合 X 中的数据进行聚类划分，然后根据划分结果再次更新人工鱼的位置，这是与经典人工鱼算法的不同之处。

3 仿真实验

仿真实验的硬件环境 CPU 为 P4 2.4GHZ，内存为 2G，软件为 Matlab 7.0。采用数据为 KDD cup-99 网络入侵检测仿真数据，KDD cup-99 数据集[10]中包含了 4 种类型的攻击，共有 22 种入侵数据。仿真实验分为算法的训练过程和测试过程，算法首先使用 KDD cup-99 的训练数据生成聚类划分的分类，然后使用 KDD cup-99 的测试数据对生成的聚类进行分类效果的测试。网络入侵检测的性能指标采用检测率和误报率[11]，检测率 DR 等于检测到的入侵个数除以训练数据集所有入侵数据个数的百分比；误报率 FAR 等于被误报为入侵的正常数据的个数除以训练数据集中正常数据个数的百分比。在网络入侵检测中，检测率 DR 越大，误报率 FAR 就越小，说明聚类的分类效果越好，反之亦然。

3.1 数据预处理

数据的特征属性分为连续型特征属性和离散型分类特征

属性，其中连续特征属性分别采用不同的度量方法，采用的度量单位越小，变量值域的变化范围就愈大；而度量单位越大，变量值域的变化范围越小，表现在计算实体之间的距离时，出现大数吃小数的现象。为了避免度量单位对聚类结果的影响，需要对连续型数值数据进行标准化、归一化的预处理；对于离散型的分类属性，根据数值的概率分布进行数值化预处理，使分类属性能够进行欧式几何距离的计算。

1)标准化处理

设 $x'_{i,j}$ 为 $x_{i,j}$ 数值标准化之后的值，计算公式如下：

$$x'_{i,j}=(x_{i,j}-mean_j)/stad_j \tag{5}$$

式中，$mean_j=\frac{1}{N}\sum_{i=1}^{N}x_{i,j}$，$stad_j=\frac{1}{N}\sum_{i=1}^{N}|x_{i,j}-mean_j|$，$mean_j$ 为第 j 个属性的均值，$stad_j$ 为第 j 个属性的平均绝对偏差。

2)归一化处理

归一化处理将数值的值域映射到[0,1]区间之内，从而避免大数吃小数的现象，使各个属性的量纲范围相同。设 $x''_{i,j}$ 为 $x'_{i,j}$ 归一化后的数值，计算公式如下：

$$x''_{i,j}=\frac{(x'_{i,j}-x'_{\min})}{(x_{\max}-x'_{\min})} \tag{6}$$

式中，$\forall i,\exists x'_{\min}\leqslant x'_{i,j},\exists x'_{i,j}\leqslant x'_{\max}$。

3)分类属性的数值化处理

分类属性需要进行数值化处理，本文采用的数值化方法是一种基于概率的方法，一个分类属性的某一个取值占的比例越大，说明该值相对于其它取值所起的作用越大，所以其数值化后的取值应越大。

设实体的第 j 个属性是离散型的分类属性，A_j 表示第 j 个属性，$A_j\in\{v_{j,1},v_{j,2},\cdots,v_{j,k}\}$，表示属性 A_j 有 k 个取值。$A_{j,s}(1\leqslant s\leqslant k)$ 映射为数值的方法如下：

$$v'_{j,s}=num(v_{j,s})/\sum_{i=1}^{k}num(v_{j,i}) \tag{7}$$

式中，$num(v_{j,s})$ 为在数据集合中 v_j 取值为 $v_{j,s}$ 的记录数量，$v'_{j,s}$ 为 $v_{j,s}$ 映射后的数值，$v'_{j,s}\in[0,1]$ 并且 $\sum_{i=1}^{k}v'_{j,s}=1$，由于 $v'_{j,s}\in[0,1]$，因此不需要再进行归一化处理。

3.2 算法测试的仿真结果

为测试本算法的分类效果，将本算法获得的聚类应用到网络入侵检测，并与经典的聚类算法 K-means 和 K-原型进行比较。由于聚类结果受聚类数目 k 的影响，因此采用不同的 k 值进行仿真实验。为了客观地反映算法的效果，减小实验误差，对于每个 k 值，做若干次仿真实验，分别计算 K-means 算法、K-原型算法和本文算法的平均检测率和平均误报率。实验结果如表 1 所列。

表 1 聚类数目与检测率和误报率之间的关系

k	K-means		K-原型		AFS-KM	
	DR/%	FAR/%	DR/%	FAR/%	DR/%	FAR/%
4	16.2	0.22	16.0	0.20	18.2	0.14
6	32.6	0.72	34.2	0.68	37.4	0.60
8	43.6	1.60	45.3	1.50	50.8	1.0
10	55.7	2.20	58.2	2.10	62.5	1.82
12	60.2	2.80	62.5	2.40	68.4	2.0
14	68.3	5.80	69.7	5.0	76.4	3.6
16	80.2	6.20	83.2	5.80	88.8	4.4
18	82.4	7.30	84.2	7.0	92.6	5.5
20	85.6	9.20	86.2	8.80	94.2	6.8

其中，DR 是算法的平均检测率，FAR 是算法的平均误报率，k 为聚类数目。由表 1 显示，AFS-KM 算法的检测率高于 K-原型和 K-means 聚类算法的检测率，而误报率低于 K-原型和 K-means 聚类算法，说明 AFS-KM 算法的分类效果优于经典的聚类算法。对实验过程跟踪分析发现，AFS-KM 算法生成的聚类与 K-means 和 K-原型相比，其聚类的类内间距较小，类内密度较大，而类间距离较大，所以分类效果较好。AFS-KM 算法进行多次随机选择训练数据，获得聚类的分类效果波动很小，而且对于同一组实验数据进行多次实验，实验结果几乎没有波动；而经典的聚类算法，对于随机选择训练数据进行聚类划分，分类效果波动较大，即使对于同样的训练数据进行多次实验，其结果也会产生一定的波动，说明经典的聚类算法对初始的聚类中心选择比较敏感，影响了实际的分类效果。另外在实验过程中发现，在没有采用加权属性进行聚类划分时，存在网络入侵的数据划分到正常数据类中的现象和正常数据划分到入侵数据类的现象，而在采用加权属相进行聚类划分时，没有发现这种现象，说明加权属性计算实体之间距离时更加科学地反映了数据的分布特性，加权属性优化了聚类的划分方法。以上分析说明，AFS-KM 算法克服了经典聚类初始中心选择敏感问题，在聚类划分时能够获得全局最优的聚类；通过加权属性能够真实地反映数据的原始分布，从而在计算欧式距离时更加真实地反映了实体之间的距离，因此优化了聚类的划分。

根据以上结果，获得 ROC 曲线，ROC 曲线是权衡检测率和误报率的一种可视化方法，用于表示误报率 FAR(False Acceptance Rate)和检测率(Detection Rate)之间关系的曲线，代表接收者的操作特性。从图 1 可以看出，检测率的增加是以误报率的增加为代价的。检测率和误报率之间存在矛盾，不能同时既获得最高的检测率，又获得最低的误报率，因而需要在它们之间确定一种折中选择的关系，如果需要获得更高的检测率，则需要牺牲一定的误报率为代价；反之，如果需要获得更低的误报率，则需要牺牲一定的检测率为代价。

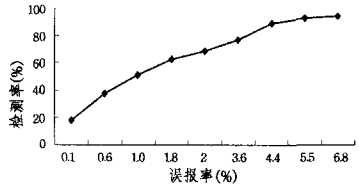

图 1 ROC 曲线

3.3 未知入侵行为检测的仿真结果

对于网络异常检测系统的要求是不但能检测到已知的入侵行为，而且能够检测到未知的网络入侵行为。为此设计了如下实验：在包含入侵标签的测试集合进行过滤，过滤部分网络入侵数据，划分出 16 个子集，每个子集包含 5000 个网络连接数据，其中每个子集分别包含 Probe、DOS、U2R 、R2L 类型的 2～200 个入侵数据，入侵数据占训练数据的比例小于等于 4%，入侵检测数据远远少于正常网络的数量，满足采用聚类算法进行网络入侵检测的假设；测试数据中除了包含训练数据的网络入侵检测数据，还包含未知的网络入侵数据，然后进行交叉验证。由表 1 显示，聚类数目为 18 时，检测率和误报率都达到理想的折中效果，因此在进行未知入侵行为检测仿真实验时聚类数目取值为 18，实验结果如表 2 所列。

表 2　不同类型未知入侵攻击行为的检测效果

入侵类型	未知入侵数据	DR/%	FAR/%	入侵个数
Probe	IP sweep	96.2	3.1	200
	Port sweep	95.2	2.0	50
	Satan	92.4	2.60	120
DOS	Back	96.3	2.63	200
	Land	100	1.20	9
	Smurf	93.5	1.84	180
	Teardrop	100	0.8	12
	Neptune	95.8	2.24	100
	Mailstorm	96.6	1.42	45
U2R	Buf_overflow	93.4	4.26	22
	Perl	94.5	2.7	20
	Sqlattack	100	1.85	85
R2L	Guess_password	94.6	1.0	165
	Multihop	100	2.22	18
	Sendmail	100	1.7	17
	Worm	100	3.2	5

由表 2 可以看出，本文提出的算法在进行入侵检测时，对于未知的网络入侵行为，也能获得较高的检测率，而同时保持较低的误报率。主要原因是：聚类算法本身是无监督的分类算法，在分类的过程中不需要已知的分类信息，如果训练模型具有较高的精度，则能获得较好的分类效果，因此本文算法适合于网络的异常检测。

3.4　算法训练的仿真结果

为了检验 AFS-KM 算法的训练时间的优劣，对 AFS-KM 算法与经典的聚类算法进行收敛时间的比较。仿真结果如图 2 所示。

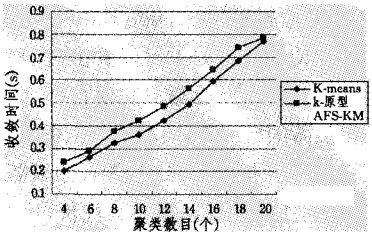

图 2　算法收敛时间的比较

图 2 表明，本文提出的 AFS-KM 算法的收敛时间明显小于 K-means 算法和 K-原型算法，原因是 AFS-KM 算法具有先天的并行性，人工鱼之间能够进行信息交流，在搜索过程中能够根据启发信息及时调整搜索方向，因此人工鱼之间通过协同工作能够快速搜索到最优的初始聚类中心。在跟踪实验过程中发现，人工鱼的随机行为能够以一定的概率快速跳出局部最优解，避免了经典聚类无启发信息的盲目搜索，所以 AFS-KM 算法收敛时间较快。

3.5　实验结果性能分析

AFS-KM 算法与经典的聚类算法 K-means、K-原型算法相比，算法的收敛时间明显减少，当聚类数目较小时，收敛时间降低较少；但随着聚类数目的增加，算法的收敛的减少时间逐渐增加。由图 2 显示，AFS-KM 算法相对于 K-原型算法，聚类收敛时间减少的最小值为 0.03s，最大值为 0.1s，其主要原因是 AFS-KM 算法中的人工鱼之间进行信息交互，能够及时更改搜索方向，从而跳出局部最优解，向全局最优解方向搜索；利用聚类算法进行网络入侵检测时，AFS-KM 与经典的聚类算法 K-means、K-原型算法相比，其网络入侵检测率、误报率都得到很大的改进；由表 1 显示，随着聚类数目的增加，检测率提高的百分比逐渐增加，误报率降低的百分比也逐渐增加，当聚类数目 $k=4$ 时，K-原型的检测率和误报率分别为 16.0%、0.20%，AFS-KM 的检测率和误报率分别为 18.2%、0.14%，此时检测率的百分比提高了 2.2%，误报率的百分比降低了 0.06%；当聚类数目 $k=20$ 时，K-原型的检测率和误报率分别为 86.2%、8.8%，AFS-KM 的检测率和误报率分别为 94.2%、6.8%，此时检测率的百分比提高了 8%，误报率的百分比降低了 2%；由表 2 显示，AFS-KM 算法进行未知入侵行为检测也能获得较好的效果，因此本文提出的 AFS-KM 算法无论是时间复杂度还是聚类的无监督分类效果，相对于经典的聚类算法都有很大的提高。

结束语　AFS-KM 算法首先利用人工鱼群算法智能搜索优化的初始聚类中心，然后运行 K-means 算法，该算法结合了人工鱼群算法和 K-means 算法的优点，既克服了 K-means 算法对于初始聚类中心选择敏感的问题，又解决了人工鱼群算法后期收敛速度较慢的问题，因而能够在相对较短的时间内获得最优的聚类划分；在聚类的过程中，计算实体之间的距离采用基于信息增益加权属性的欧式距离计算方法，加权属性欧式距离充分体现了各个属性在分类过程中贡献的大小，真实地反映了数据的原始分布特征。实验结果表明本文算法应用于网络入侵检测时，能够获得理想的网络入侵检测率和误报率。

参 考 文 献

[1] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data [M]. China: Posts & Telecom Press, 2009

[2] 梁吉业，白亮，等. 基于新的距离度量的 K-modes 聚类算法[J]. 计算机研究与发展，2010，47(10)：1749-1755

[3] 胡艳维，秦拯，张忠志. 基于模拟退火与 K 均值聚类的入侵检测算法[J]. 计算机科学，2010，6(6)：122-124

[4] Das S, Abraham A, Konar A. Automatic clustering using an improved differential evolution algorithm[J]. IEEE Trans. on System, Man cybernetics-part: System, Humans, 2008, 38(1): 218-237

[5] Bandyopadhyay S, Saha S. A point symmetry-based clustering technique for automatic evolution of clusters[J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(11): 1441-1457

[6] Xie Zong-bo, Feng Jiu-chao. A Sparse Projection clutering Algorithm[J]. Journal of Electronics, 2009, 26(4): 549-551

[7] 公茂果，王爽，等. 复杂分布数据的二阶段聚类算法[J]. 软件学报，2011，22(11)：2760-2772

[8] Theodoridis S. 模式识别[M]. 北京：电子工业出版社，2010

[9] 李晓磊，邵之江，钱积新. 一种基于动物自治体的寻优模式：鱼群算法[J]. 系统工程理论与实践，2002，22(11)：32-38

[10] KDD Cup 1999 Data. University of California, Irivine[OL]. http://kdd.ics.uci.edu/databases/kddcup99/ kddcup99.1999

[11] 肖立中，邵志清，马汉华，等. 网络入侵检测中自动决定聚类数算法[J]. 软件学报，2008，19(8)：2140-2148