权 利 要 求 书

1. 一种网络连接数据分类系统，用于对 D 个不同的网络连接数据进行分类，其特征在于，包括：

数据存储部；分类设定部；粒子随机生成部；距离计算部；数据分类部；判断设定部；位置和变化率调整部；分类结束判断部；以及结果输出部，

其中，所述数据存储部存储有所述 D 个网络连接数据，每个所述网络连接数据含有 d 个特征属性值，

所述分类设定部设定 m 个分类，

所述粒子随机生成部生成 n 个空间粒子，每个所述空间粒子在与所述 d 个特征属性值相对应的 d 维求解空间中的当前位置为（P_{d1}，P_{d2}，•••，P_{dm}），每个所述空间粒子的当前变化率为（v_{d1}，v_{d2}，•••，v_{dm}），每个中心数据 P_{di}（i=1，•••，m）包含与所述 d 个特征属性值相对应的 d 个粒子位置属性值，每个 v_{di} 包含与所述中心数据 P_{di} 的所述 d 个粒子位置属性值相对应的 d 个中心粒子变化率，

所述距离计算部分别计算每个所述网络连接数据与每个所述空间粒子的 m 个中心数据 P_{di} 之间的距离，

所述数据分类部根据每个所述网络连接数据与每个所述空间粒子的 m 个中心数据 P_{di} 之间的所述距离的大小将所有所述网络连接数据分成 m 类，所述数据分类部根据 n 个所述空间粒子对所述网络连接数据进行 n 次分类，

所述距离计算部计算每次分类中的所有所述网络连接数据到对应的中心数据 P_{di} 之间的距离和，

所述判断设定部判断 n 次分类中所述距离和最小的空间粒子，并设定该空间粒子为基准粒子，

所述位置和变化率调整部调整除所述基准粒子外的其他空间粒子的当前位置以及当前变化率，

所述距离计算部再次分别计算每个所述网络连接数据与每个所述粒子的 m 个中心数据 P_{di} 之间的距离，

所述数据分类部再次根据每个所述网络连接数据与每个所述粒子的 m 个中心数据 P_{di} 之间的所述距离的大小将所有所述网络连接数据分成 m 类，所述数据分类部再次根据 n 个所述空间粒子对所述网络连接数据进行 n 次分类，

所述距离计算部再次计算每次所有所述网络连接数据到对应的中心数据 P_{di} 之间的距离和，

所述分类结束判断部判断所述位置和变化率调整部调整的次数是否大于预定次数，并判断相邻两次调整的距离和差值是否小于预定阈值，

当两个判断中任意一个为是时，所述结果输出部将所述基准粒子作为分类中心，m 个中心数据 P_{di} 所在的分类作为最终分类进行输出，

当判断均为否时，所述位置和变化率调整部再次调整除所述基准粒子外的粒子的当前位置以及当前变化率，

n、m、D、d 均为大于 1 的正整数，且 D>m。

2. 根据权利要求 1 所述的网络连接数据分类系统，其特征在于：

其中，d 个所述特征属性值含有连接时间、连接的数据包、网络服务类型、连接标记以及连接时的记录参数。

3. 根据权利要求 1 所述的网络连接数据分类系统，其特征在于，还包括：

存储控制部，

其中，所述存储控制部控制所述数据存储部存储所述最终分类。

4. 根据权利要求 1 所述的网络连接数据分类系统，其特征在于：

其中，所述位置和变化率调整部包括：

位置变化率调整单元，根据每个所述其他空间粒子的所述当前变化率调整每个所述其他空间粒子的当前位置，并根据所述基准粒子的当前变化率调整所述其他空间粒子的当前变化率。

5. 根据权利要求 4 所述的网络连接数据分类系统，其特征在于：

其中，所述位置和变化率调整部还包括：

第一交叉位置变化率生成单元，在所述位置变化率调整单元调整所述其他空间粒子的所述当前位置和所述当前变化率后，选取 n 个粒子中任意 Z 个粒子并将 Z 个粒子中任意两个粒子的当前位置以及当前变化率进行交叉运算生成第一交叉位置以及第一交叉变化率，

$5\%\times n \le Z \le 40\%\times n$，Z 为正整数。

6. 根据权利要求 5 所述的网络连接数据分类系统，其特征在于：

其中，所述位置和变化率调整部还包括：

父本选择单元，选择所述当前基准粒子作为父本；

变化率位置叠加单元，选取进行交叉运算后的 n 个粒子中任意 k 个粒子，并将所述父本的当前位置以及当前变化率与被选取的所述粒子的当前位置以及当前变化率分别叠加；

第二交叉位置变化率生成单元，将叠加后的所有粒子不重复地两两配对，并再次执行交叉运算生成第二交叉位置以及第二交叉变化率；以及

位置变化率变异单元，对每个生成的第二交叉位置以及第二交叉变化率的空间粒子进行变异运算重新生成作为变异位置的当前位置以及作为变异变化率的当前变化率，

$5\%\times n \le k \le 40\%\times n$，k 为正整数。

7.根据权利要求 5 或 6 所述的网络连接数据分类系统，其特征在于：

其中，所述交叉运算的运算公式如下：

$$\begin{cases} \hat{x}_1^{iter} = p \cdot x_1^{iter} + (1-p) \cdot x_2^{iter} \\ \hat{x}_2^{iter} = p \cdot x_2^{iter} + (1-p) \cdot x_1^{iter} \\ \hat{v}_1^{iter} = p \cdot v_1^{iter} + (1-p) \cdot v_2^{iter} \\ \hat{v}_2^{iter} = p \cdot v_2^{iter} + (1-p) \cdot v_1^{iter} \end{cases}$$

其中，*iter* 代表当前生成位置和变化率的调整次数，x_1，x_2，v_1，v_2 分别代表选择交叉运算前的两个粒子的当前位置和当前变化率，$\hat{x}_1$，

$\hat{x}_2$，$\hat{v}_1$，$\hat{v}_2$ 分别代表选择交叉运算后的两个粒子的当前位置和当前变化率。

8.根据权利要求 5 所述的网络连接数据分类系统，其特征在于：

其中，所述变异运算的运算公式如下：

$$x_k^{iter+1}=\begin{cases}\hat{x}_k^{iter}+c_k & if \quad fit(\hat{x}_k^{iter}+c_k)>fit(\hat{x}_k^{iter}) \quad and \quad r>0.5\\ \hat{x}_k^{iter} & otherwise\end{cases}$$

$$\hat{v}_k^{iter+1}=\begin{cases}0.5\cdot(\hat{v}_k^{iter}+v_k^{iter}) & if \quad r<0.5\\ \hat{v}_k^{iter} & otherwise\end{cases},$$

c_k 是区间 $\left[x^L-\hat{x}_k^{iter},x^U-\hat{x}_k^{iter}\right]$ 上均匀分布的随机数，x^L，x^U 分别是可行区间的边界，*fit* 代表适应度函数。

网络连接数据分类系统

技术领域

本发明具体涉及一种网络连接数据分类系统。

背景技术

随着近年来互联网的爆炸式普及，网络已经深入人们的生活、娱乐和工作中。但互联网的开放性和安全性是一把双刃剑，它在给人们带来便利的同时，互联网的无主管性、跨国性、不设防性使得网络安全问题越来越突出。网络入侵检测是网络安全系统的重要组成部分，其对未经授权的使用、滥用网络资源的行为进行监控和响应，具有保护信息完整性、机密性作用。

通常来说，网络入侵检测方法包括异常入侵检测和误用入侵检测方法。误用入侵检测方法认为异常行为和正常行为之间的交集很大，其检测结果与检测知识库完备性密切相关，不能发现新入侵行为，检测结果没有实际意义，因此异常入侵检测方法是当前网络入侵检测主要研究方向。异常检测是以网络的正常运行状态为基础，构造模型及规则来描述网络在正常情况下的各种特征。将当前网络特征发生较大偏差时来判断网络是否有异常或攻击存在。

数据挖掘是异常入侵检测系统中当前最主要的网络入侵检测工具。数据挖掘主要对互联网的网络纪录进行分析，从中挖掘隐含的、实现未知的潜在有用信息，并用这些信息去检测异常入侵和已知的入

侵。

为了保证数据挖掘的准确率并减小误警率，需要事先构建准确的网络连接数据的分类，但是在构建数据分类的过程中，往往容易陷入局部最优的问题，造成分类相当不准确。

发明内容

本发明是为了解决上述问题而进行的，目的在于提供一种提高网络连接数据的分类准确率的网络连接数据分类系统。

本发明提供了一种网络连接数据分类系统，用于对 D 个不同的网络连接数据进行分类，具有这样的特征，包括：数据存储部；分类设定部；粒子随机生成部；距离计算部；数据分类部；判断设定部；位置和变化率调整部；分类结束判断部；以及结果输出部，其中，数据存储部存储有 D 个网络连接数据，每个网络连接数据含有 d 个特征属性值，分类设定部设定 m 个分类，粒子随机生成部生成 n 个粒子，每个粒子的当前位置为（P_{d1}，P_{d2}，…，P_{dm}），每个粒子的当前变化率为（v_{d1}，v_{d2}，…，v_{dm}），每个中心数据 P_{di}（i=1，…，m）包含与 d 个特征属性值相对应的 d 个粒子位置属性值，每个 v_{di} 包含与中心数据 P_{di} 的 d 个粒子位置属性值相对应的 d 个中心粒子变化率，距离计算部分别计算每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离，数据分类部根据每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离的大小将所有网络连接数据分成 m 类，数据分类部对网络连接数据进行 n 次分类，距离计算部计算每次分类中

的所有网络连接数据到对应的中心数据 P_{di} 之间的距离和，判断设定部判断 n 次分类中距离和最小的粒子，并设定该粒子为基准粒子，位置和变化率调整部调整除基准粒子外的其他空间粒子的当前位置以及当前变化率，距离计算部再次分别计算每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离，数据分类部再次根据每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离的大小将所有网络连接数据分成 m 类，数据分类部再次对网络连接数据进行 n 次分类，距离计算部再次计算每次所有网络连接数据到对应的中心数据 P_{di} 之间的距离和，分类结束判断部判断位置和变化率调整部调整的次数是否大于到预定次数，并判断相邻两次调整的距离和差值是否小于预定阈值，当两个判断中任意一个为是时，结果输出部将基准粒子作为分类中心，m 个中心数据 P_{di} 所在的分类作为最终分类进行输出，当判断均为否时，位置和变化率调整部再次调整除基准粒子外的粒子的当前位置以及当前变化率，n、m、D、d 均为大于 1 的正整数，且 D>m。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，d 个特征属性值含有连接时间、连接的数据包、网络服务类型、连接标记以及连接时的记录参数。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征，还包括：存储控制部，其中，存储控制部控制数据存储部存储最终分类。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，位置和变化率调整部包括：位置变化率调整单元，根据每

个其他空间粒子的当前变化率调整每个其他空间粒子的当前位置，并根据基准粒子的当前变化率调整其他空间粒子的当前变化率。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，位置和变化率调整部还包括：第一交叉位置变化率生成单元，在位置变化率调整单元调整其他空间粒子的当前位置和当前变化率后，选取 n 个粒子中任意 Z 个粒子并将 Z 个粒子中任意两个粒子的当前位置以及当前变化率进行交叉运算生成第一交叉位置以及第一交叉变化率，$5\%\times n \le Z \le 40\%\times n$，Z 为正整数。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，位置和变化率调整部还包括：父本选择单元，选择当前基准粒子作为父本；变化率位置叠加单元，选取进行交叉运算后的 n 个粒子中任意 k 个粒子，并将父本的当前位置以及当前变化率与被选取的粒子的当前位置以及当前变化率分别叠加；第二交叉位置变化率生成单元，将叠加后的所有粒子不重复地两两配对，并再次执行交叉运算生成第二交叉位置以及第二交叉变化率；以及变化率变异单元，对每个生成的第二交叉位置以及第二交叉变化率的空间粒子进行变异运算重新生成作为变异位置的当前位置以及作为变异变化率的当前变化率，$5\%\times n \le k \le 14\%\times n$，k 为正整数。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，交叉运算的运算公式如下：

$$\begin{cases} \hat{x}_1^{iter} = p \cdot x_1^{iter} + (1-p) \cdot x_2^{iter} \\ \hat{x}_2^{iter} = p \cdot x_2^{iter} + (1-p) \cdot x_1^{iter} \\ \hat{v}_1^{iter} = p \cdot v_1^{iter} + (1-p) \cdot v_2^{iter} \\ \hat{v}_2^{iter} = p \cdot v_2^{iter} + (1-p) \cdot v_1^{iter} \end{cases}$$

其中，*iter* 代表当前生成位置和变化率的调整次数，x_1，x_2，v_1，v_2 分别代表选择交叉运算前的两个粒子的当前位置和当前变化率，$\hat{x}_1$，$\hat{x}_2$，$\hat{v}_1$，$\hat{v}_2$ 分别代表选择交叉运算后的两个粒子的当前位置和当前变化率。

在本发明提供的网络连接数据分类系统中，还可以具有这样的特征：其中，变异运算的运算公式如下：

$$x_k^{iter+1} = \begin{cases} \hat{x}_k^{iter} + c_k & if \quad fit(\hat{x}_k^{iter} + c_k) > fit(\hat{x}_k^{iter}) \quad and \quad r > 0.5 \\ \hat{x}_k^{iter} & otherwise \end{cases}$$

$$\hat{v}_k^{iter+1} = \begin{cases} 0.5 \cdot (\hat{v}_k^{iter} + v_k^{iter}) & if \quad r < 0.5 \\ \hat{v}_k^{iter} & otherwise \end{cases},$$

c_k 是区间 $\left[x^L - \hat{x}_k^{iter}, x^U - \hat{x}_k^{iter}\right]$ 上均匀分布的随机数，x^L，x^U 分别是可行区间的边界，*fit* 代表适应度函数。

发明的作用与效果

根据本发明所涉及的网络连接数据分类系统，因为具有数据存储部；分类设定部；粒子随机生成部；距离计算部；数据分类部；判断设定部；位置和变化率调整部；分类结束判断部；以及结果输出部，所以，本发明的网络连接数据分类系统可以准确地对网络连接数据进行分类，而且具有更高的检测率和更低的误报率，且具有较好的收敛

性。不仅如此，本发明的网络连接数据分类系统还可以用于对运营数据的异常数据、证券交易数据的异常数据进行准确分类，并有效检测判断出异常数据。

附图说明

图 1 是本发明的实施例中网络连接数据分类系统的框图；

图 2 是本发明的实施例中网络连接数据分类系统的动作流程图；

图 3 是本发明的实施例中位置和变化率调整部的动作流程图；以及

图 4 是本发明的实施例中网络连接数据分类系统的分类效果图。

具体实施方式

为了使本发明实现的技术手段、创作特征、达成目的与功效易于明白了解，以下实施例结合附图对本发明网络连接数据分类系统作具体阐述。

图 1 是本发明的实施例中网络连接数据分类系统的框图。

如图 1 所示，网络连接数据分类系统 10 具有数据存储部 11、分类设定部 12、粒子随机生成部 13、距离计算部 14、数据分类部 15、判断设定部 16、位置和变化率调整部 17、分类结束判断部 18、结果输出部 19、存储控制部 20 以及控制部 21。

数据存储部11存储有一个网络流量测试数据集，在本实施例中，该网络流量测试数据集为KDD Cup 99数据集，KDD Cup 99数据集是由麻省理工学院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，且该数据集包含了7个星期网络流量，大约50万条网络连接数据，考虑到KDD Cup 99数据集比较庞大，所以将其分为训练集A1和测试集A2，其中训练集A1用来生成检测模型，主要是用来生成分类需要的分类中心向量；余下的数据作为进行验证的测试集A2，（即、D=25万）。这些网络连接数据中包括多种广泛的网络环境下的模拟入侵，包括22种攻击类型和1个正常类型，如下表1所示。

表 1 网络连接数据标识类型

标识类型	含义	具体分类标识
Normal	正常记录	normal
DoS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probe	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对超级权限的非法访问	buffer_overflow、loadmodule、perl、rootkit

从上表可以看出网络连接数据集中的异常类型按攻击手段分为：DoS、R2L、U2R、Probe四类。其中每个连接实例包含42个属性且均标识为正常或特定的攻击类型。数据集的数据格式如下：

0, udp, private, SF，105，146，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，0，2，2，0.00，0.00，0.00，0.00，1.00，0.00，0.00，255, 254，1.00，0.01，0.00，0.00，0.00，0.00，0.00，0.00，snmpgetattack

在这条数据中，第一个属性是连接时间；第二个属性表明该连接是 TCP 还是 UDP 数据包；第三个属性表示服务类型，如 http、ftp、smtp 等；第四个属性表明连接标记，如 SF、REJ、RSTR 等；随后的 37 个为该连接的数值属性，即连接时的记录参数；最后一个属性是

类标记属性，表明这条记录是正常连接还是入侵连接。在本实施例中，d 为 41，在 41 个固定的特征属性中，9 个特征属性为离散(symbolic)型,其他均为连续(continuous)型。

分类设定部 12 设定分类的数目，在本实施例中，数目为 m 个，m 为大于 1 的正整数。

粒子随机生成部 13 生成 n 个粒子，n 为大于 1 的正整数。每个粒子的当前位置为（P_{d1}，P_{d2}，…，P_{dm}），每个粒子的当前变化率为（v_{d1}，v_{d2}，…，v_{dm}），每个中心数据 P_{di}（i=1，…，m）包含与 d 个特征属性值相对应的 d 个粒子位置属性值，每个 v_{di} 包含与中心数据 P_{di} 的 d 个粒子位置属性值相对应的 d 个中心粒子变化率。

距离计算部 14 根据每个网络连接数据的前 41 个特征属性值分别计算每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离，并计算每次分类中的所有网络连接数据到对应的中心数据 P_{di} 之间的距离和。

当某个特征能使不同类别的网络连接数据之间具有最大距离，而同类网络连接数据之间具有最小距离时，算法赋予该特征最高Fisher分值。当d=2时，粒子的当前位置以及当前变化率均符合

$$X=\{(x_1,y_1),(x_2,y_2),\bullet\bullet\bullet,(x_m,y_m)\},\ x_i(i=1,2,\bullet\bullet\bullet,D)\in R^d,$$

d为特征空间的维数，类标记为 $y_i\in\{+1,-1\}$，D为网络连接数据数。如此Fisher分值定义为：

$$F=S_b/S_w$$

其中 S_b 表示类间离散度和，描述两类网络连接数据间的距离；

S_w为类内离散度和，描述同类网络连接数据间的离散度和。定义 $S_b=(\overline{m}_1-\overline{m})^2+(\overline{m}_2-\overline{m})^2$， $\overline{m}_1,\overline{m}_2,\overline{m}$分别为正常网络连接数据、异常网络连接数据和所有网络连接数据的均值。由此可以得到 $S_w=S_p+S_n=\sigma_p^2+\sigma_n^2$， σ_p^2,σ_n^2分别为正常网络连接数据、异常网络连接数据的方差。对于数据集中的41个属性可以得到第r个属性的Fisher分值表达式为

$$F_r=\frac{S_b}{S_w}=\sum_{i=1}^{2}(\overline{m}_{i,r}-\overline{m}_r)^2/\sum_{i=1}^{2}\sigma_{i,r}^2$$

同理，式中$\overline{m}_{i,r},\overline{m}_r$分别为第$i$类网络连接数据和所有网络连接数据的第$r$个特征的均值； $\sigma_{i,r}^2$为第i类网络连接数据第r个特征的方差。计算41个属性的Fisher分值可以得到其排序。

数据分类部15根据每个网络连接数据与每个粒子的m个中心数据P_{di}之间的距离的大小将所有网络连接数据分成m类。数据分类部15根据n个粒子对网络连接数据进行n次分类。

判断设定部16判断n次分类中距离和最小的粒子，并设定该粒子为基准粒子。

位置和变化率调整部17调整除基准粒子外的其他空间粒子的当前位置以及当前变化率。

位置和变化率调整部17包括：位置变化率调整单元171、第一交叉位置变化率生成单元172、父本选择单元173、变化率位置叠加单元174、第二交叉位置变化率生成单元175以及位置变化率变异单元176。

位置变化率调整单元171根据每个其他空间粒子的当前变化率

调整每个其他空间粒子的当前位置，并根据基准粒子的当前变化率调整其他空间粒子的当前变化率。

网络连接数据的当前变化率、当前位置的调整方程表示为：

$$v_{id}^{k+1}=\omega\cdot v_{id}^{k}+c_1\xi\left(p_{id}^{k}-x_{id}^{k}\right)+c_2\eta\left(p_{gd}^{k}-x_{id}^{k}\right)$$
$$x_{id}^{k+1}=x_{id}^{k}+\gamma\cdot v_{id}^{k+1}$$

在网络连接数据集中每个网络连接数据都是 d 维空间内的一个点。第 i 个网络连接数据可以表示为自身搜索到的历史最优值 p_i，$p_i=(p_{i1},p_{i2},...,p_{iD})$，$p_g$ 为所有网络连接数据搜索到的最优值，c_i 是网络连接数据跟踪自己历史最优值的权重系数，它表示网络连接数据自身的认识。c_2 是网络连接数据跟踪群体最优值的权重系数，它表示网络连接数据对整个群体知识的认识。ξ，η 是[0,1]区间内均匀分布的随机数。γ 是对位置更新变化率系数。

ω 是保持当前变化率的系数，表示网络连接数据的先前变化率对当前变化率的影响程度。若 ω 较大，网络连接数据有能力扩展搜索空间，全局搜索能力强。若 ω 较小，网络连接数据主要在当前粒子的附近搜索，局部搜索能力较强。改变 ω 的取值可以调整算法全局和局部搜索能力。ω 由式子：$\omega=\omega_{\max}-(\omega_{\max}-\omega_{\min})/iter_{\max}\times iter$ 确定，其中 $iter_{max}$ 是调整次数的最大值，$iter$ 是当前调整次数。

在位置变化率调整单元 171 调整其他空间粒子的当前位置和当前变化率后，第一交叉位置变化率生成单元 172 选取 n 个粒子中任意 Z 个粒子并将 Z 个粒子中任意两个不重复的粒子的当前位置以及当前变化率进行交叉运算生成第一交叉位置以及第一交叉变化率。Z 的取

值范围是 5%×n ≤ Z ≤ 40%×n，且 Z 为正整数。

交叉运算的运算公式如下：

$$\begin{cases} \hat{x}_1^{iter} = p \cdot x_1^{iter} + (1-p) \cdot x_2^{iter} \\ \hat{x}_2^{iter} = p \cdot x_2^{iter} + (1-p) \cdot x_1^{iter} \\ \hat{v}_1^{iter} = p \cdot v_1^{iter} + (1-p) \cdot v_2^{iter} \\ \hat{v}_2^{iter} = p \cdot v_2^{iter} + (1-p) \cdot v_1^{iter} \end{cases}$$

其中，*iter* 代表当前生成位置和变化率的调整次数，x_1，x_2，v_1，v_2 分别代表选择交叉运算前的两个粒子的当前位置和当前变化率，$\hat{x}_1$，$\hat{x}_2$，$\hat{v}_1$，$\hat{v}_2$ 分别代表选择交叉运算后的两个粒子的当前位置和当前变化率。

父本选择单元 173 选择当前基准粒子作为父本。

变化率位置叠加单元 174 选取进行交叉运算后的 n 个粒子中任意 k 个粒子，并将父本的当前位置以及当前变化率与被选取的粒子的当前位置以及当前变化率分别叠加，k 的取值范围为 5%×n ≤ k ≤ 14%×n，k 为正整数。

第二交叉位置变化率生成单元 175 将叠加后的所有粒子不重复地两两配对，并再次执行交叉运算生成第二交叉位置以及第二交叉变化率。

位置变化率变异单元 176 对每个生成的第二交叉位置以及第二交叉变化率的空间粒子进行变异运算重新生成作为变异位置的当前位置以及作为变异变化率的当前变化率。

变异运算的运算公式如下：

$$x_k^{iter+1}=\begin{cases}\hat{x}_k^{iter}+c_k & if \quad fit(\hat{x}_k^{iter}+c_k)>fit(\hat{x}_k^{iter}) \quad and \quad r>0.5\\ \hat{x}_k^{iter} & otherwise\end{cases}$$

$$\hat{v}_k^{iter+1}=\begin{cases}0.5\cdot(\hat{v}_k^{iter}+v_k^{iter}) & if \quad r<0.5\\ \hat{v}_k^{iter} & otherwise\end{cases},$$

c_k 是区间 $\left[x^L-\hat{x}_k^{iter},x^U-\hat{x}_k^{iter}\right]$ 上均匀分布的随机数，x^L，x^U 分别是可行区间的边界，*fit* 代表适应度函数。

距离计算部 14 再次分别计算每个网络连接数据与每个调整后的粒子的 m 个中心数据 P_{di} 之间的距离。

分类结束判断部 18 判断位置和变化率调整部 17 调整的次数是否大于到预定次数，并判断相邻两次调整的距离和差值是否小于预定阈值。在本实施例中，预定次数为 400 次，预定阈值为万分之一。

当两个判断中任意一个为是时，结果输出部 19 将基准粒子作为分类中心，m 个中心数据 P_{di} 所在的分类作为最终分类进行输出。结果输出部 19 对不同的类别设定不同的编号。结果输出部 19 给定分类中心后分类的划分按照最邻近法则决定：

若对于某一个网络连接数据 X_i 和分类编号 j 若满足：

$$\left\|X_i-Z_j\right\|=\min_{k=1,2,\cdots,m}\left\|X_i-Z_k\right\|$$

则说明该网络连接数据取到所有分类的最佳匹配，X_i 属于第 j 类。

当两个判断均为否时，位置和变化率调整部 17 再次调整除基准粒子外的粒子的当前位置以及当前变化率。

存储控制部 20 控制数据存储部 11 存储最终分类。

控制部 21 包含用于控制数据存储部 11、分类设定部 12、粒子随

机生成部 13、距离计算部 14、数据分类部 15、判断设定部 16、位置和变化率调整部 17、分类结束判断部 18、结果输出部 19 以及存储控制部 20 运行的计算机程序。

图 2 是本发明的实施例中网络连接数据分类系统的动作流程图。

如图 2 所示，本实施例的网络连接数据分类系统 10 的动作流程图包含以下步骤：

步骤 S1-1，分类设定部 12 设定 m 个分类，然后进入步骤 S1-2。

步骤 S1-2，粒子随机生成部 13 生成 n 个粒子，然后进入步骤 S1-3。

步骤 S1-3，距离计算部 14 分别计算每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离，然后进入步骤 S1-4。

步骤 S1-4，数据分类部 15 根据每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离的大小将所有网络连接数据分成 m 类，然后进入步骤 S1-5。

步骤 S1-5，距离计算部 14 计算每次分类中的所有网络连接数据到对应的中心数据 P_{di} 之间的距离和，然后进入步骤 S1-6。

步骤 S1-6，判断设定部 16 判断 n 次分类中距离和最小的粒子，并设定该粒子为基准粒子，然后进入步骤 S1-7。

步骤 S1-7，位置和变化率调整部 17 调整除基准粒子外的其他空间粒子的当前位置以及当前变化率，然后进入步骤 S1-8。

步骤 S1-8，距离计算部 14 再次分别计算每个网络连接数据与每个粒子的 m 个中心数据 P_{di} 之间的距离，然后进入步骤 S1-9。

步骤 S1-9，数据分类部 15 再次根据每个网络连接数据与每个粒

子的 m 个中心数据 P_{di} 之间的距离的大小将所有网络连接数据分成 m 类，然后进入步骤 S1-10。

步骤 S1-10，距离计算部 14 再次计算每次分类中的所有网络连接数据到对应的中心数据 P_{di} 之间的距离和，然后进入步骤 S1-11。

步骤 S1-11，分类结束判断部 18 判断位置和变化率调整部调整的次数是否大于预定次数，并判断相邻两次调整的距离和差值是否小于预定阈值，当判断均为否时，进入步骤 S1-7；当两个判断中任意一个为是时，进入步骤 S1-12。

步骤 S1-12，结果输出部 19 将基准粒子作为分类中心，m 个中心数据 P_{di} 所在的分类作为最终分类进行输出，然后进入步骤 S1-13。

步骤 S1-13，存储控制部 20 控制数据存储部存储最终分类，然后进入结束状态。

图 3 是本发明的实施例中位置和变化率调整部的动作流程图。

如图 3 所示，本实施例的位置和变化率调整部 17 的动作流程图包含以下步骤：

步骤 S7-1，位置变化率调整单元 171 根据每个其他空间粒子的当前变化率调整每个其他空间粒子的当前位置，并根据基准粒子的当前变化率调整其他空间粒子的当前变化率，然后进入步骤 S7-2。

步骤 S7-2，第一交叉位置变化率生成单元 172 选取 n 个粒子中任意 Z 个粒子并将 Z 个粒子中任意两个不重复的粒子的当前位置以及当前变化率进行交叉运算生成第一交叉位置以及第一交叉变化率，然后进入步骤 S7-3。

步骤 S7-3，父本选择单元 173 选择当前基准粒子作为父本，然后进入步骤 S7-4。

步骤 S7-4，变化率位置叠加单元 174 选取进行交叉运算后的 n 个粒子中任意 k 个粒子，并将父本的当前位置以及当前变化率与被选取的粒子的当前位置以及当前变化率分别叠加，然后进入步骤 S7-5。

步骤 S7-5，第二交叉位置变化率生成单元 175 将叠加后的所有粒子不重复地两两配对，并再次执行交叉运算生成第二交叉位置以及第二交叉变化率，然后进入步骤 S7-6。

步骤 S7-6，位置变化率变异单元 176 对每个生成的第二交叉位置以及第二交叉变化率的空间粒子进行变异运算重新生成作为变异位置的当前位置以及作为变异变化率的当前变化率，然后进入结束状态。

实验结果对比及分析

实验的数据集选取了比较权威 KDD Cup 99 数据的“kddcup.data_10.percent”，该数据集共有 491421 条记录，正常的总和为 97278 条，其余的 396473 均为异常型。异常的分为四类：DoS、U2R、R2L 和 Probe。其中每种类型的具体标识种类在表 1 中列出。在“kddcup.data_10.percent”数据集中被识别出来的标识有 22 种攻击类型。为了评价分析结果，采用误报率 FAR 和检测率 DR 来衡量。其定义描述如下：

FAR=被误判为入侵的正常记录数/总测试记录中的正常记录数；

DR=检测出来的入侵记录数/总测试中的入侵记录数。

分类算法能够应用在网络异常检测是基于以下两个基本的假设：

(1)正常数据的数量远远大于异常数据量；

(2)异常数据在某些属性的取值上明显偏离正常的取值范围。

实验环境：本实验的软硬件环境为：CPU：主频3.0GHz，内存4GB，操作系统Windows7以及Matlab2014a。配置主要参数为：分类数目 m=2；粒子种群规模D=15；最大调整次数 $\max\_iter$=400；交叉、变异概率 p_c, p_m=rand[0,1]； c_1, c_2 均为1。

从测试集 A2 提取出 4 组作为测试样本。详细列出随机抽取的各个样本的集合如下表 2 所示。

表 2 数据集选取和分类表

数据集类别	正常数据	异常数据	正常比例
训练集 A1	19152	498	97%
测试集 A2-DoS	24256	744	97%
测试集 A2-R2L	24256	744	97%
测试集 A2-Probe	24256	744	97%
测试集 A2-混合	24256	744	97%

该数据集随机抽样满足上述分类算法应用在异常检测中的数据抽取要求，可以作为实验数据进行后续分析。由于该数据集中属性特征之间存在着很大差异性，而且它们可能是采用不同的单位来度量。为了消除由于度量标准的不同对分类的影响，必须对样本中的数据做标准归一化处理，即将原始数据从原来所处空间转换到一个标准化空间。对于一个 $n\times k$ 的矩阵，方法如下：

$$\hat{x}_{ij} = \left(x_{ij} - \overline{x_j}\right) / S_j, \quad i = 1,2,\cdots,n; j = 1,2,\cdots,k$$

其中，$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$， $S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}$ 即为标准化后的实验数据

值。通过计算每个特征值与平均值之间的标准偏差，可得到该特征值存正规化空间中的新值。

试验开始需要先获取分类中心，选取训练集 A1 进行普通 K 均值分类,将该结果保存作为后续使用。

Fisher 分值定义为：$F = S_b / S_w$，其中 S_b 表示类间离散度，描述两类样本间的距离；S_w 为类内离散度，描述同类样本间的离散度。定义 $S_b = (\overline{m}_1 - \overline{m})^2 + (\overline{m}_2 - \overline{m})^2$，$\overline{m}_1, \overline{m}_2, \overline{m}$ 分别为正常样本、异常样本和所有样本的均值。由此可以得到 $S_w = S_p + S_n = \sigma_p^2 + \sigma_n^2$，$\sigma_p^2, \sigma_n^2$ 分别为正常样本，异常样本的方差。对于数据集中的 41 个属性可以得到第 r 个属性的 Fisher 分值表达式为 $F_r = \frac{S_b}{S_w} = \sum_{i=1}^{2} (\overline{m}_{i,r} - \overline{m}_r)^2 / \sum_{i=1}^{2} \sigma_{i,r}^2$。式中 $\overline{m}_{i,r}, \overline{m}_r$ 分别为第 i 类样本和所有样本的第 r 个特征的均值；$\sigma_{i,r}^2$ 为第 i 类样本第 r 个特征的方差。计算 41 个属性的 Fisher 分值可以得到其排序。在进行 Fisher 分排序时不用具体区分攻击方式，即将所有入侵类型都归为异常，这样形成了二值分类问题。按照 Fisher 判别法得到 41 个属性 Fisher 分排序为：

(12,23,32,2,24,36,31,6,39,25,26,38,29,4,34,33,37,35,13,28,27,41,14,3,19,8,13,22,14,18,7,11,5,15,1,17,16,10,9,20,21)。

为了验证该 Fisher 排序进行特征提取的作用，设计实验，将排序的结果抽取前 10，20，13 分别自成一组特征组，随机抽取 10，20，13 个特征分别自成一组特征组，将 41 个属性全部作为一组特征组，分别对这 7 个特征组输入测试集 A2-混合类型测试，采用本实施例中的网络连接数据分类系统统计 FAR，DR 和运行时间如下表 3 所示。

表 3 网络连接数据分类系统的特征提取列表

分组类型	FAR	DR	运行时间
Fisher 排序前 10	8.3%	78.2%	54s
Fisher 排序前 20	7.4%	81.5%	88s
Fisher 排序前 13	7.2%	85.1%	132s
随机抽取 10	10.2%	17.6%	16s
随机抽取 20	9.7%	72.8%	93s
完全属性 41	12.2%	64.3%	203s

从上述表格中可以看出 Fisher 排序提取特征属性能够极大地减少运行时间。可以看出异常检测的误报率在 Fisher 排序后相对于随机抽取和完全属性组有改善，说明有些冗余特征属性会给异常检测带来干扰。

在上述实验的基础上，本发明选取 Fisher 排序前 15 个特征作为该 PSO 算法的输入数据向量，并比较位置和变化率调整部 17 中仅采用位置变化率调整单元 171 的网络连接数据分类系统（第一分类）、仅采用位置变化率调整单元 171 和第一交叉位置变化率生成单元 172（第二分类）的网络连接数据分类系统以及采用位置变化率调整单元 171、第一交叉位置变化率生成单元 172、父本选择单元 173、变化率位置叠加单元 174、第二交叉位置变化率生成单元 175 和位置变化率变异单元 176（第三分类）的网络连接数据分类系统的性能。如下表 4 给出 3 种算法的在测试集 A2 中的检测结果和运行时间。

表 4 三种算法检测效果对比表

组编号	第一分类			第二分类			第三分类		
	DR	FAR	Time	DR	FAR	Time	DR	FAR	Time
A2-DoS	68.6%	18.5%	75s	84.3%	8.2%	95s	94.9%	2.5%	108s
A2-R2L	64.8%	19.3%	79s	18.1%	10.4	94s	95.3%	3.1%	100s
A2-Probe	67.4%	21.2%	76s	79.8%	11.7%	89s	94.2%	4.7%	102s

A2-混合	65.3%	24.7%	71s	82.4%	9.8%	94s	93.7%	3.6%	101s

从上表看出采用第一分类的装置异常检测效果明显低于第三分类的网络连接数据分类系统，而采用第二分类的装置要略优于采用第一分类的装置。当然，在时间消耗上采用第三分类的网络连接数据分类系统相对其他两种来说较多。

图 4 是本发明的实施例中网络连接数据分类系统的分类效果图。

如图 4 所示，本实施例的网络连接数据分类系统 10 在采用第三分类的分类过程中后期收敛，前期有略微波动。虽然第三分类在第 261 次后出现跳动，是由于本实施例在研究过程中加入了变异因子，其虽然引起了短期内的跳动，但为后代的持续优化提供了更好的基础，因此此处跳动属于增加变异因子的正常现象。而第一分类的收敛变化率最快，也很容易陷入局部最优值；第二分类的收敛过程较为平稳，但是最终的离散度和整体高于第三分类的网络连接数据分类系统。

实施例的作用与效果

根据本实施例所涉及的网络连接数据分类系统，因为具有数据存储部；分类设定部；粒子随机生成部；距离计算部；数据分类部；判断设定部；位置和变化率调整部；分类结束判断部；以及结果输出部，所以，本实施例的网络连接数据分类系统可以准确地对网络连接数据进行分类，而且具有更高的检测率和更低的误报率，且具有较好的收敛性。不仅如此，本实施例的网络连接数据分类系统还可以用于对运营数据的异常数据、证券交易数据的异常数据进行准确分类，并有效

检测判断出异常数据。

上述实施方式为本发明的优选案例，并不用来限制本发明的保护范围。

5

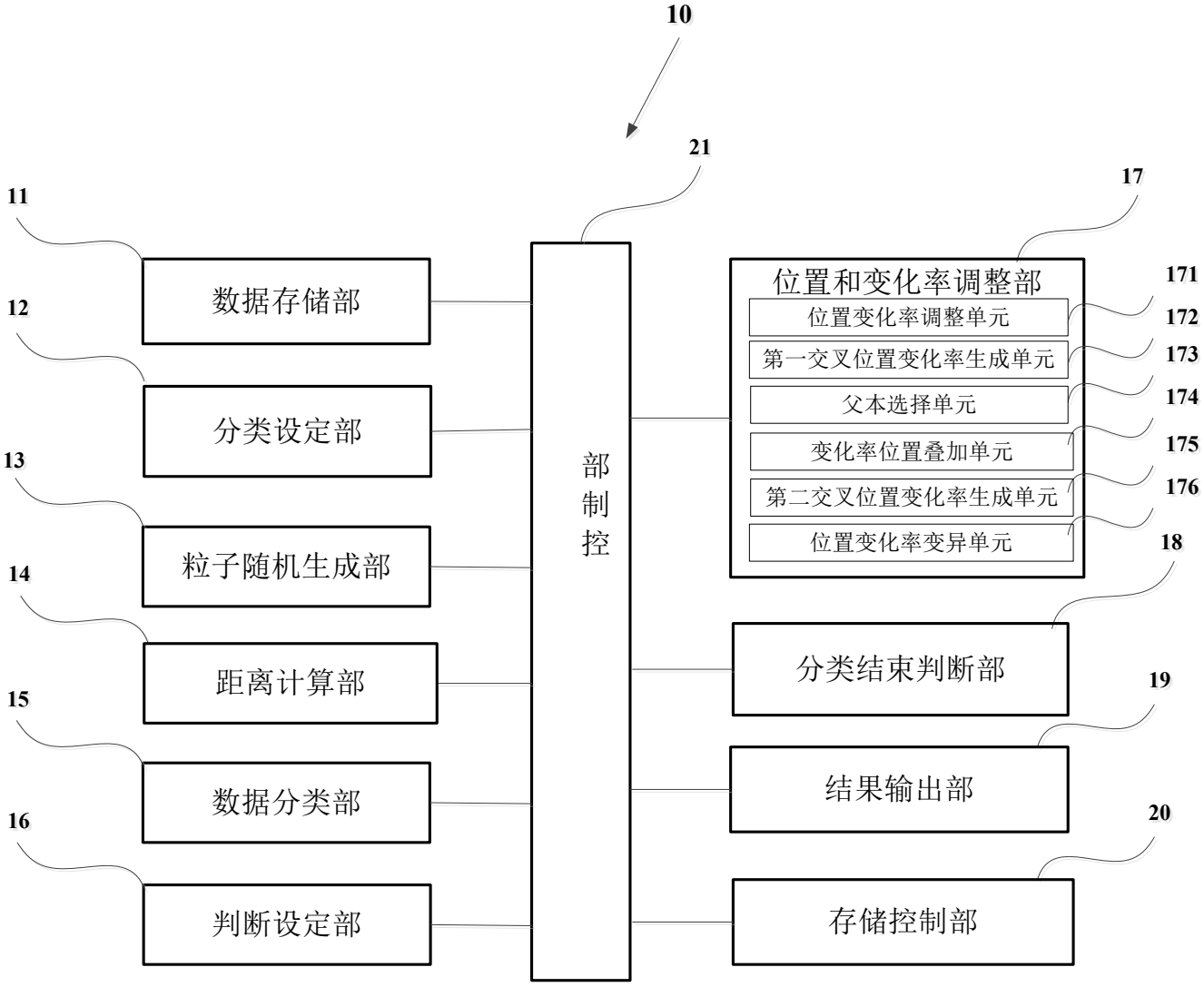

10
21
11
数据存储部
12
分类设定部
13
粒子随机生成部
14
距离计算部
15
数据分类部
16
判断设定部
部
制
控
17
位置和变化率调整部
171
位置变化率调整单元
172
第一交叉位置变化率生成单元
173
父本选择单元
174
变化率位置叠加单元
175
第二交叉位置变化率生成单元
176
位置变化率变异单元
18
分类结束判断部
19
结果输出部
20
存储控制部

图 1

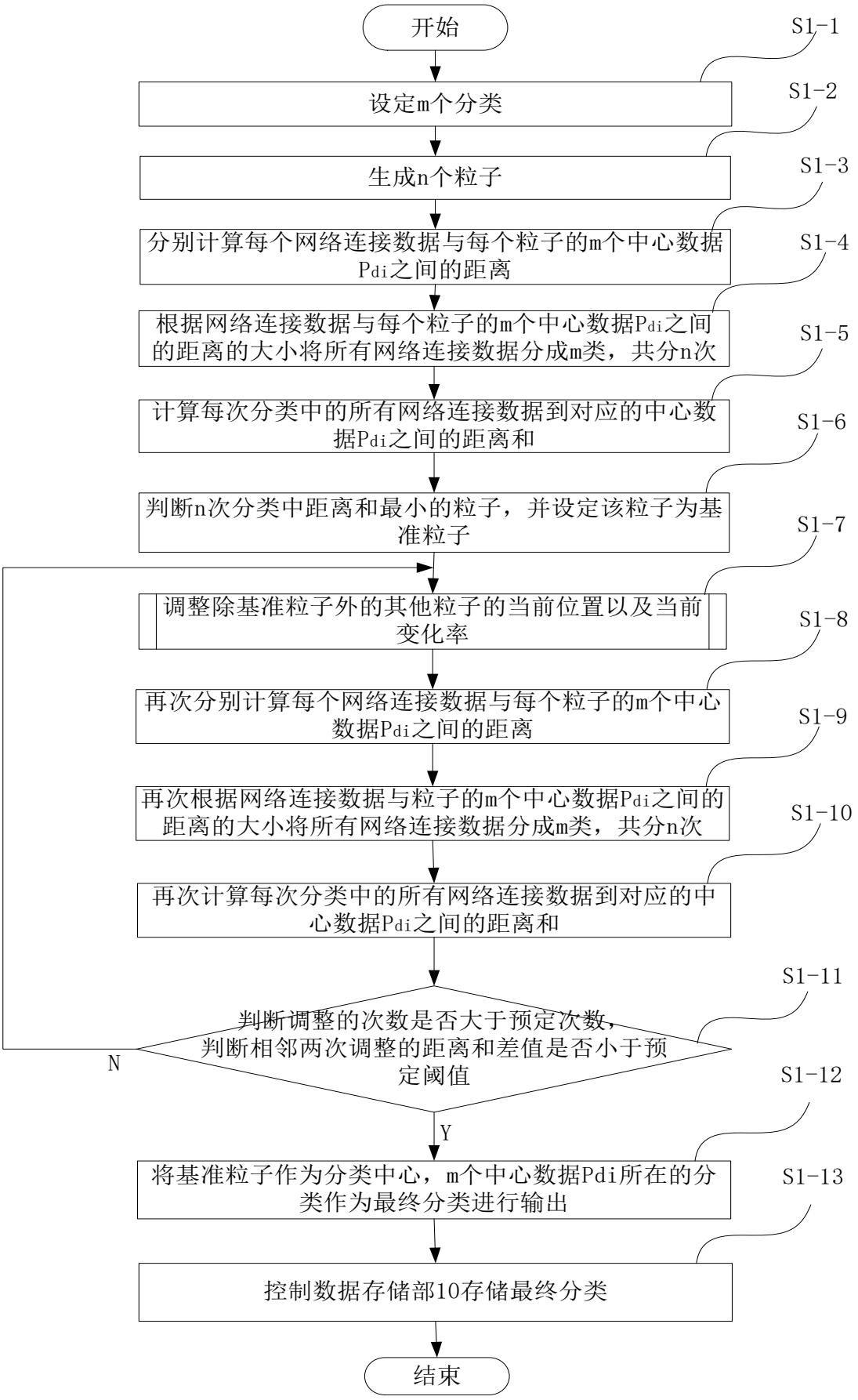

开始
S1-1
设定m个分类
S1-2
生成n个粒子
S1-3
分别计算每个网络连接数据与每个粒子的m个中心数据Pdi之间的距离
S1-4
根据网络连接数据与每个粒子的m个中心数据Pdi之间的距离的大小将所有网络连接数据分成m类，共分n次
S1-5
计算每次分类中的所有网络连接数据到对应的中心数据Pdi之间的距离和
S1-6
判断n次分类中距离和最小的粒子，并设定该粒子为基准粒子
S1-7
调整除基准粒子外的其他粒子的当前位置以及当前变化率
S1-8
再次分别计算每个网络连接数据与每个粒子的m个中心数据Pdi之间的距离
S1-9
再次根据网络连接数据与粒子的m个中心数据Pdi之间的距离的大小将所有网络连接数据分成m类，共分n次
S1-10
再次计算每次分类中的所有网络连接数据到对应的中心数据Pdi之间的距离和
S1-11
判断调整的次数是否大于预定次数，判断相邻两次调整的距离和差值是否小于预定阈值
N
Y
S1-12
将基准粒子作为分类中心，m个中心数据Pdi所在的分类作为最终分类进行输出
S1-13
控制数据存储部10存储最终分类
结束

图 2

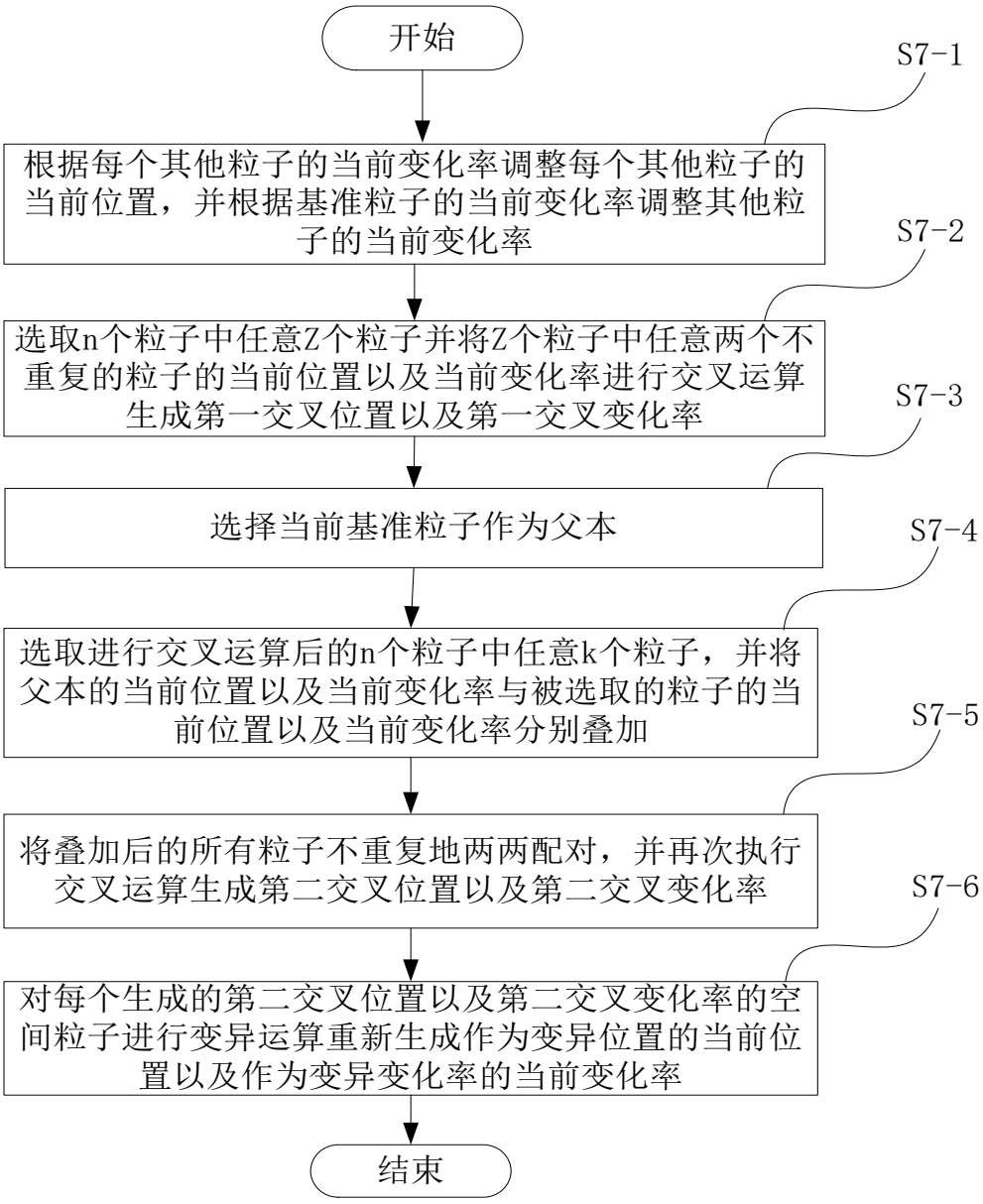

开始
S7-1
根据每个其他粒子的当前变化率调整每个其他粒子的当前位置，并根据基准粒子的当前变化率调整其他粒子的当前变化率
S7-2
选取n个粒子中任意Z个粒子并将Z个粒子中任意两个不重复的粒子的当前位置以及当前变化率进行交叉运算生成第一交叉位置以及第一交叉变化率
S7-3
选择当前基准粒子作为父本
S7-4
选取进行交叉运算后的n个粒子中任意k个粒子，并将父本的当前位置以及当前变化率与被选取的粒子的当前位置以及当前变化率分别叠加
S7-5
将叠加后的所有粒子不重复地两两配对，并再次执行交叉运算生成第二交叉位置以及第二交叉变化率
S7-6
对每个生成的第二交叉位置以及第二交叉变化率的空间粒子进行变异运算重新生成作为变异位置的当前位置以及作为变异变化率的当前变化率
结束

图 3

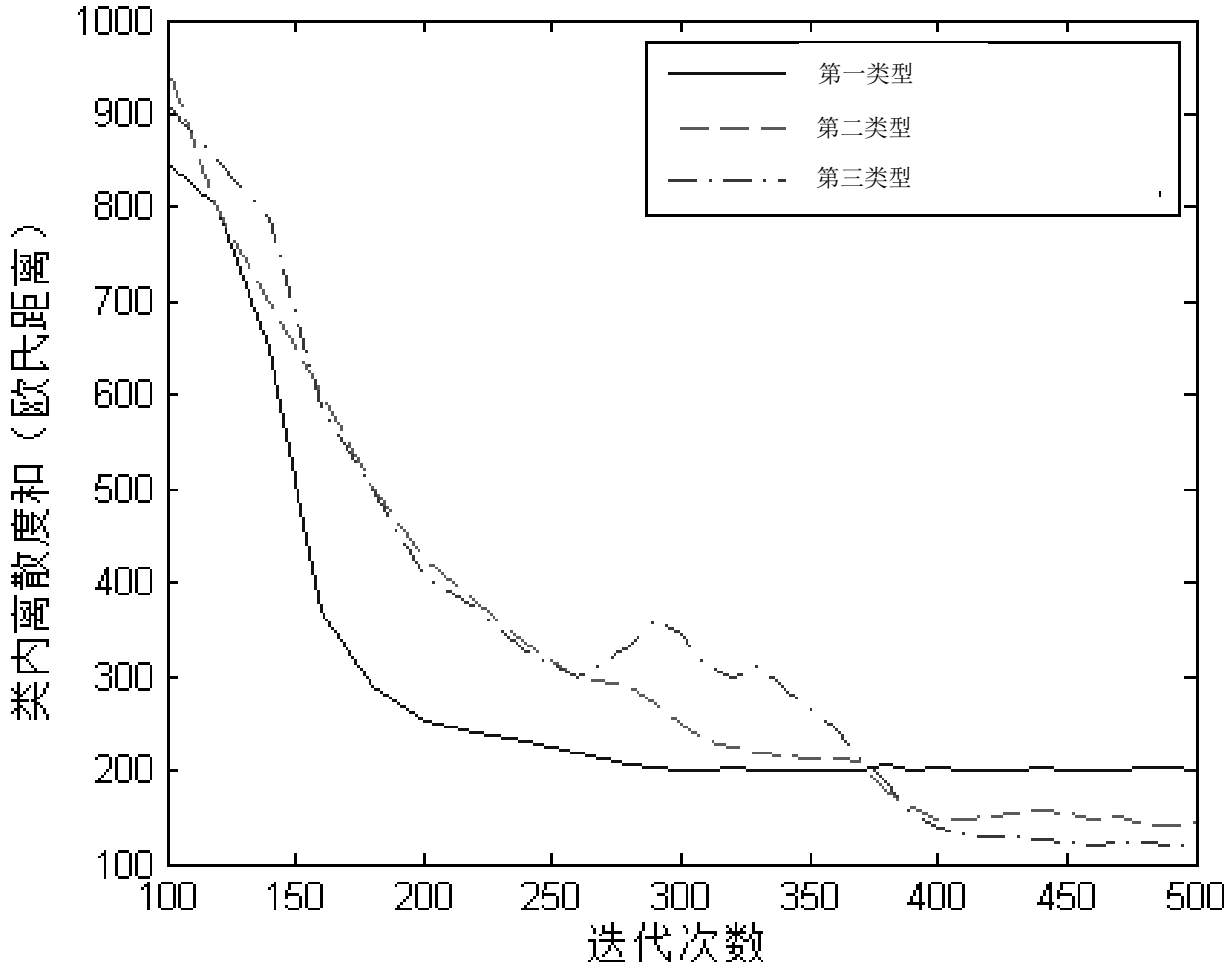
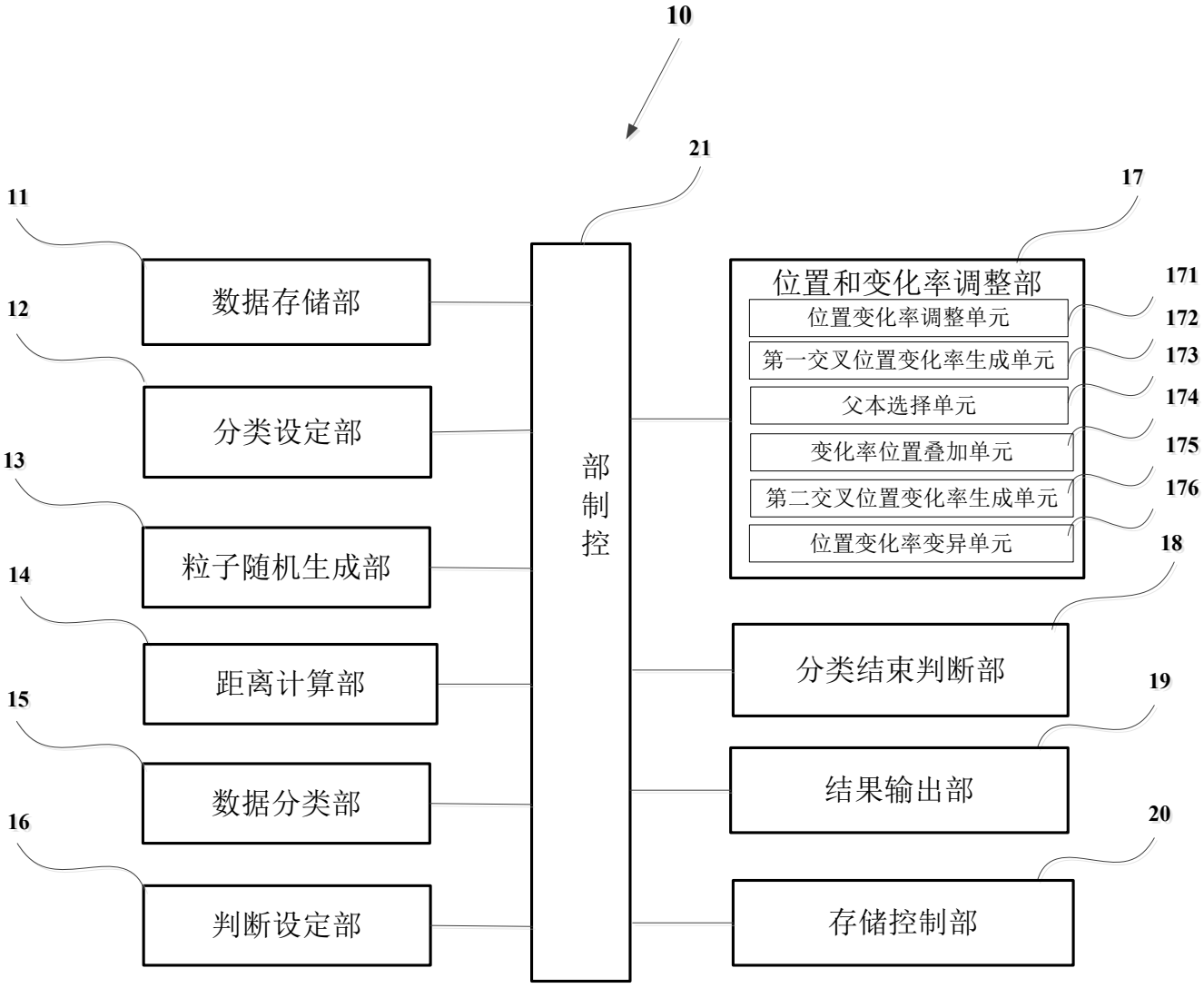

1000
900
800
700
600
500
400
300
200
100
100
150
200
250
300
350
400
450
500
类内离散度和（欧氏距离）
迭代次数
第一类型
第二类型
第三类型

图 4

说 明 书 摘 要

本发明提供了一种提高网络连接数据的分类准确率的网络连接数据分类装置。本发明提供的网络连接数据分类装置，用于对 D 个不同的网络连接数据进行分类，包括：数据存储部；分类设定部；粒子随机生成部；距离计算部；数据分类部；判断设定部；位置和变化率调整部；分类结束判断部；以及结果输出部，其中，数据存储部存储有 D 个网络连接数据，分类设定部设定 m 个分类，粒子随机生成部生成 n 个粒子，距离计算部分别计算距离，数据分类部根据距离的大小将所有网络连接数据分成 m 类，判断设定部判断 n 次分类中距离和最小的粒子，位置和变化率调整部调整当前位置以及当前变化率，分类结束判断部判断是否结束分类，结果输出部将最终分类进行输出。

摘 要 附 图


10
21
11
数据存储部
12
分类设定部
13
粒子随机生成部
14
距离计算部
15
数据分类部
16
判断设定部
部
制
控
17
位置和变化率调整部
171
位置变化率调整单元
172
第一交叉位置变化率生成单元
173
父本选择单元
174
变化率位置叠加单元
175
第二交叉位置变化率生成单元
176
位置变化率变异单元
18
分类结束判断部
19
结果输出部
20
存储控制部