(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109218321 A
(43)申请公布日 2019.01.15

(21)申请号 201811116207.X

(22)申请日 2018.09.25

(71)申请人 北京明朝万达科技股份有限公司
地址 100097 北京市海淀区蓝靛厂南路25号嘉友国际大厦北区2层

(72)发明人 曾毅 孙加光 喻波 王志海 董爱华 安鹏

(51)Int.Cl.
H04L 29/06(2006.01)
H04L 12/24(2006.01)

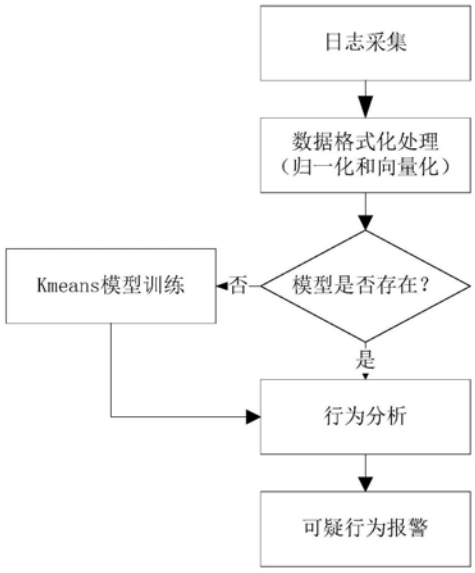
权利要求书2页 说明书6页 附图3页

(54)发明名称

一种网络入侵检测方法及系统

(57)摘要

本发明提供了一种网络入侵检测方法,包括以下步骤:(1)对历史终端行为数据进行数据处理;(2)判断是否存在Kmeans聚类分析模型,如果不存在,根据历史终端行为数据建立基于spark的Kmeans聚类分析模型,进入步骤(3),如果存在,直接进行步骤(3);(3)根据所述Kmeans聚类分析模型对新接收到的终端行为数据进行分析;(4)检测到网络入侵和可疑连接行为,并进行攻击链地图展示。提高功能扩展和时效性,降低漏报率,使人机界面更为友好。



CN 109218321 A

1.一种网络入侵检测方法,包括以下步骤:(1)对历史终端行为数据进行数据处理;(2)判断是否存在Kmeans聚类分析模型,如果不存在,根据历史终端行为数据建立基于spark的Kmeans聚类分析模型,进入步骤(3),如果存在,直接进行步骤(3);(3)根据所述Kmeans聚类分析模型对新接收到的终端行为数据进行分析;(4)检测到网络入侵和可疑连接行为,并进行攻击链地图展示。

2.如权利要求1所述的方法,其中步骤(1)的终端行为数据采用日志方式来记录每一次网络连接访问的情况,日志中的主要字段包括:源IP地址、目的IP地址、目的端口号、请求接口、操作类型、请求发生的时间、操作持续时间,所述日志能表征访问的行为特征,可用该数据来标识终端访问的属性。

3.如权利要求1-2任一项所述的方法,对访问服务器的终端行为数据进行采集,并对采集的数据进行归一化处理和向量化处理,将每条终端行为记录转化为一个与请求相关的终端行为向量。

4.如权利要求1-3任一项所述的方法,其中步骤(2)建立基于spark的Kmeans聚类分析模型为,以历史终端行为向量作为输入数据对模型进行训练,得到收敛的聚类结果,将训练好的聚类分析模型保存下来,对聚类分析模型进行训练的步骤如下:步骤a:选择聚类的个数k并对训练向量集中的数据进行归一化和向量化处理;步骤b:随机选取训练向量集中的k个向量作为聚类中心;步骤c:计算训练向量集中每个向量与k个聚类中心的距离,并将该向量分配给与其距离最近的聚类中心,从而获得k个簇群;步骤d:对每个簇群计算该簇群所有数据点的平均值,并将其作为该簇群的新聚类中心;步骤e:重复步骤c-d,直到聚类中心不再改变,由此获得k个簇群;

其中k≥1。

5.如权利要求1-4任一项所述的方法,步骤(3)为:对新采集到的终端行为数据,计算其与各聚类分析中心的距离,当距离超过该阀值时,判断为该新采集到的终端行为数据为可疑行为数据。

6.如权利要求1-5任一项所述的方法,步骤(4)具体为,对步骤(3)中发现的可疑行为数据进行分析,找到可疑行为数据对应的源ip地址和目的ip地址,将该源ip地址和目的ip地址与对应的位置信息关联,以攻击链的形式展示在地图上。

7.一种网络入侵检测系统,包括:

数据处理单元,用于对终端行为数据进行数据处理;

判断单元,判断是否存在Kmeans聚类分析模型;

聚类分析模型建立单元,用于根据历史终端行为数据建立基于spark的Kmeans聚类分析模型;

数据分析单元,用于对新接收到的终端行为数据进行分析;

异常行为分析结果展示单元,用于检测到网络入侵和可疑连接行为时,进行攻击链地图展示。

8.如权利要求7所述的系统,其中数据处理单元用于通过数据采集模块对访问服务器的终端行为数据进行采集,并对采集的数据进行归一化处理和向量化处理,将每条终端行为记录转化为一个与请求相关的终端行为向量。

9.如权利要求7-8任一项所述的系统,其中聚类分析模型建立单元用于以历史终端行

为向量作为输入数据，通过大量训练数据对聚类分析模型进行训练，得到收敛的聚类结果，保存训练好的聚类分析模型。

10.如权利要求7-9任一项所述的系统，数据分析单元用于对终端行为数据进行分析，具体为对新采集到的终端行为数据向量，计算其与各聚类分析中心的距离，设定一个阀值，当距离超过该阀值时，判断为该终端行为数据为可疑行为数据。

11.如权利要求7-10任一项所述的系统，异常行为分析结果展示单元对数据分析单元中发现的可疑数据进行分析，找到可疑行为数据对应的源ip地址和目的ip地址，将该源ip地址和目的ip地址与对应的位置信息关联，以攻击链的形式展示在地图上。

一种网络入侵检测方法及系统

技术领域

[0001] 本发明属于计算机网络技术领域,涉及一种终端异常行为分析方法。

背景技术

[0002] 分析终端异常行为、检测网络入侵是要找到与以往所见的正常连接不同的连接,对保障网络安全至关重要。现有常用的检测网络入侵的方法为基于异常的网络入侵检测系统(A-NIDS)。A-NIDS又分为基于异常数据包的入侵检测和基于异常网络流量的入侵检测。基于异常数据包的入侵检测中,从网络中获得数据来源,根据异常数据包特征库的特征,对捕获到的数据包进行实时分析,若数据包与异常数据包特征库中某一特征相匹配,则认为是一个异常数据包。基于异常数据包的入侵检测的缺点是需要不断更新异常数据包特征库,否则如果异常数据包的特征未收集在异常数据包特征库,则难以识别异常数据包。基于异常网络流量的入侵检测中,根据采样所得到的样本来预测当前网络的流量,若当前网络流量与预测相比有较大差异,具体表现为突发性网络流量激增,可以认为网络流量发生异常。当异常确定为入侵时,将产生报警。基于异常网络流量的入侵检测的缺点是小流量下的入侵不能被检测到,且如果正常流量行为突然出现增加,会导致系统误判。

[0003] 聚类就是对大量未知标注的数据集,按数据的内在相似性将数据集划分为多个类别,使类别内的数据相似度较大而类别间的数据相似度较小。聚类是最有名的非监督学习算法,Kmeans是应用最广泛的聚类算法,它试图在数据集中找出k个簇群。在Kmeans算法中,数据点是由所有数值型特征组成的特征向量,简称向量。数据点相互距离一般采用欧氏距离,如点1(x11,x12,x13,…,x1n),点2(x21,x22,x23,…,x2n)之间的距离计算如下:

[0004] $$D_{12}=\sqrt{\sum_{i=1}^{n}(x_{1i}-x_{2i})^2}$$

[0005] Kmeans算法中簇群实质上是一个点,即组成该簇群的所有点的中心(称为质心),它是簇群中所有点的算术平均值,因此算法取名Kmeans。算法开始时随机选择k个数据点作为簇群的质心,然后把每个数据点分配给最近的质心,接着对每个簇群计算该簇群所有数据点的平均值,并将其作为该簇群的新质心,然后不断重复该过程直到质心稳定不再变化,由此得到k个簇群。以数据集合为三维,簇群内为两点,两点分别为X=(x1,x2,x3),Y=(y1,y2,y3)为例,中心点Z变为Z=(z1,z2,z3),其中z1=(x1+y1)/2,z2=(x2+y2)/2,z3=(x3+y3)/2。Kmeans聚类算法有以下优点:1.作为解决聚类问题的一种经典算法,具有简单、快速的特点;2.对处理大数据集具有可伸缩性和高效率;3.当结果簇密集时效果较好。虽然目前已有将Kmeans模型用于网络入侵检测的方法(如CN 107895171A),但该方法通过深度置信网络与Kmeans算法结合分类,算法复杂,且Kmeans模型的训练数据为几种攻击类型的异常数据,在异常数据收集不完整时存在出现漏报的风险。

发明内容

[0006] 针对现有入侵检测方法中存在的问题,基于Kmeans算法特点,本发明提供了一种基于spark的机器学习Kmeans分析模型的终端异常行为分析方法。检测网络入侵及可疑连

接的本质是要找到与以往见过的正常连接不同的连接，本发明根据每个网络连接的统计属性进行聚类，聚类的结果簇定义了正常的历史连接类型，界定了正常连接的区域，任何在区域之外的点都是不正常的、可疑的，将正常连接区域之外的点认定为网络入侵。本发明还将终端行为分析结果的攻击链通过地图显示，并提供了相应的分析系统。

[0007] 为解决上述技术问题，本发明一实施例提供了一种网络入侵检测方法，包括以下步骤：(1)对历史终端行为数据进行数据处理；(2)判断是否存在Kmeans聚类分析模型，如果不存在，根据历史终端行为数据建立基于spark的Kmeans聚类分析模型，进入步骤(3)，如果存在，直接进行步骤(3)；(3)根据所述Kmeans聚类分析模型对新接收到的终端行为数据进行分析；(4)检测到网络入侵和可疑连接行为，并进行攻击链地图展示。

[0008] 根据本发明的方法，优选的，其中步骤(1)的终端行为数据采用日志方式来记录每一次网络连接访问的情况，日志中的主要字段包括：源IP地址、目的IP地址、目的端口号、请求接口、操作类型、请求发生的时间、操作持续时间，所述日志能表征访问的行为特征，可用该数据来标识终端访问的属性。

[0009] 根据本发明的方法，优选的，对访问服务器的终端行为数据进行采集，并对采集的数据进行归一化处理和向量化处理，将每条终端行为记录转化为一个与请求相关的终端行为向量。

[0010] 根据本发明的方法，优选的，其中步骤(2)建立基于spark的Kmeans聚类分析模型为，以历史终端行为向量作为输入数据对模型进行训练，得到收敛的聚类结果，将训练好的聚类分析模型保存下来，对聚类分析模型进行训练的步骤如下：步骤a：选择聚类的个数k并对训练向量集中的数据进行归一化和向量化处理；步骤b:随机选取训练向量集中的k个向量作为聚类中心；步骤c：计算训练向量集中每个向量与k个聚类中心的距离，并将该向量分配给与其距离最近的聚类中心，从而获得k个簇群；步骤d:对每个簇群计算该簇群所有数据点的平均值，并将其作为该簇群的新聚类中心；步骤e:重复步骤c-d，直到聚类中心不再改变，由此获得k个簇群；

[0011] 其中$k \geq 1$。

[0012] 根据本发明的方法，优选的，步骤(3)为：对新采集到的终端行为数据，计算其与各聚类分析中心的距离，当距离超过该阀值时，判断为该新采集到的终端行为数据为可疑行为数据。

[0013] 根据本发明的方法，优选的，步骤(4)具体为，对步骤(3)中发现的可疑行为数据进行分析，找到可疑行为数据对应的源ip地址和目的ip地址，将该源ip地址和目的ip地址与对应的位置信息关联，以攻击链的形式展示在地图上。

[0014] 为解决上述技术问题，本发明又一实施例提供了一种网络入侵检测系统，包括：

[0015] 数据处理单元，用于对终端行为数据进行数据处理；

[0016] 判断单元，判断是否存在Kmeans聚类分析模型；

[0017] 聚类分析模型建立单元，用于根据历史终端行为数据建立基于spark的Kmeans聚类分析模型；

[0018] 数据分析单元，用于对新接收到的终端行为数据进行分析；

[0019] 异常行为分析结果展示单元，用于检测到网络入侵和可疑连接行为时，进行攻击链地图展示。

[0020] 根据本发明的系统，优选的，其中数据处理单元用于通过数据采集模块对访问服务器的终端行为数据进行采集，并对采集的数据进行归一化处理和向量化处理，将每条终端行为记录转化为一个与请求相关的终端行为向量。

[0021] 根据本发明的系统，优选的，其中聚类分析模型建立单元用于以历史终端行为向量作为输入数据，通过大量训练数据对聚类分析模型进行训练，得到收敛的聚类结果，保存训练好的聚类分析模型。

[0022] 根据本发明的系统，优选的，数据分析单元用于对终端行为数据进行分析，具体为对新采集到的终端行为数据向量，计算其与各聚类分析中心的距离，设定一个阀值，当距离超过该阀值时，判断为该终端行为数据为可疑行为数据。

[0023] 根据本发明的系统，优选的，异常行为分析结果展示单元对数据分析单元中发现的可疑数据进行分析，找到可疑行为数据对应的源ip地址和目的ip地址，将该源ip地址和目的ip地址与对应的位置信息关联，以攻击链的形式展示在地图上。

[0024] 本发明取得了如下有益效果：

[0025] 1.功能扩展和时效性：基于终端行为流量数据的机器学习威胁检测方法可以快速发现网络中的异常行为数据，使得数据的分析以近乎实时的速度完成，及时向用户报警，提高处理威胁发现处理效率，强化了系统审计功能和警报功能的时效性。

[0026] 2.漏报率低：对历史终端数据建模，与聚类中心距离阈值大于设定值即为异常数据，能够大大降低漏报率。

[0027] 3.人机界面友好性：以直观的方式将整个攻击链按ip地址的位置信息展示在地图上，以便使用者快速定位到问题的所在。

[0028] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在所写的说明书、权利要求书、以及附图中所特别指出的结构来实现和获得。

附图说明

[0029] 此处所说明的附图用来提供对本发明的进一步理解，构成本发明的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

[0030] 图1为本发明的网络入侵检测流程方法流程图；

[0031] 图2为本发明的网络入侵检测流程系统组成图；

[0032] 图3为基于spark的Kmeans聚类分析模型的建立过程示意图。

具体实施方式

[0033] 以下结合说明书附图对本发明的优选实施例进行说明，应当理解，此处所描述的优选实施例仅用于说明和解释本发明，并不用于限定本发明，并且在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互组合。

[0034] 检测网络入侵及可疑连接的本质是要找到与以往见过的连接不同的连接。可根据每个网络连接的统计属性进行聚类，结果簇定义了历史连接类型，帮我们界定了正常连接的区域。任何在区域之外的点都是不正常的，可疑的。

[0035] 终端行为采用日志方式来记录每一次访问的情况，日志中主要字段为：源IP地址、

目的IP地址、目的端口号、请求接口、操作类型、请求发生的时间、操作持续时间等。输出的日志能很好的表征访问的行为特征，可用该数据来标识终端访问的属性。

[0036] Kmeans试图在数据集中找出k个簇群，数据集中的点其实就是由所有数值型特征组成的特征向量，简称向量。Kmeans算法中簇群其实是一个点，即组成该簇的所有点的中心。例如：数据集合为三维，聚类以两点：X＝(x1,x2,x3)，Y＝(y1,y2,y3)。中心点Z变为Z＝(z1,z2,z3)，其中z1＝(x1+y1)/2，z2＝(x2+y2)/2，z3＝(x3+y3)/2。

[0037] 实施例1

[0038] Kmeans分析模型的建立及行为分析判断过程如图1所示，包括以下步骤：

[0039] (1)对终端历史行为数据进行数据处理的过程

[0040] 终端历史行为采用日志方式来记录每一次访问的情况，日志中主要字段为：源IP地址、目的IP地址、目的端口号、请求接口、操作类型、请求发生的时间、操作持续时间等。对采集的终端历史行为数据进行归一化处理，处理后的消息字段如下：

[0041] type TerminalAction struct{

[0042] var treceived string//服务器收到终端请求的时间

[0043] var duration int//操作持续时间

[0044] val sip string//终端ip地址

[0045] val dip string//终端请求的ip地址

[0046] val dport string//终端请求的端口号

[0047] val interface string//接口名

[0048] var action int//操作类型(0:增、1:删、2:查、3:改)

[0049] val reqLen int64//请求包的长度(字节数)

[0050] val resLen int64//响应包的长度(字节数)

[0051] (2)对于归一化处理后的终端历史行为数据进行向量化处理

[0052] 将每条终端历史行为记录(TerminalAction)转化为一个与请求相关的vector(向量)，

[0053] (terminalId,deviceId,dport,action,interface,treceived,duration,resLen,reqLen)。

[0054] 向量的具体创建规则如下：

[0055] TerminalId(终端编号)

[0056] 根据数据中的sip字段关联出对应的终端id号(用于标识该sip的唯一数字编号)。该编号为自定义编号，只要用于唯一标识一台终端或者服务器即可。

[0057] DestinationIp(目的ip)

[0058] 根据数据中的dip字段关联出对应服务器id号(用于标识该dip的唯一数字编号)。该编号为自定义编号，只要用于唯一标识一台终端或者服务器即可。

[0059] DestinationPort(目的端口)

[0060] 使用数据中的dport字段。

[0061] Time of day(时间)

[0062] 使用数据中的treceived字段。对应产生操作时时间的小时数值。

[0063] Request Bytes(请求参数的大小)

[0064] 使用数据中的resLen字段对应值所对应的相应区间的编号。如下所示[0,512,1024,2048,4096,…],单位为字节数。即0-512对应1,512-1024对应2,以此类推,如果resLen等于256字节,则对应的值为1;如果resLen等于760字节,则对应的值为2。

[0065] Response Bytes(响应结果的大小)

[0066] 使用数据中的reqLen字段对应值所对应的相应区间的编号。如下所示[0,512,1024,2048,4096,…],单位为字节数。即0-512对应1,512-1024对应2,以此类推,如果resLen等于256字节,则对应的值为1;如果resLen等于760字节,则对应的值为2。

[0067] Interface(操作类型)

[0068] 访问的接口名对应的编码。接口与编码之间的关系根据实际情况定义。如

[0069] /szga/login对应的编码为0001;/terminals/create对应的编码为0002。.

[0070] Action(操作类型)

[0071] 0对应增加;1对应删除;2对应查询;3对应修改。

[0072] Duration(操作持续时间)

[0073] 整个操作从请求到响应的时间对应所在区间的编号,如下所示[0,10,20,30,40,50,60,70,…],单位为秒,即0-10对应1,10-20对应2,以此类推,如果duration等于10秒,则对应的值为2。

[0074] 对于一条归一化后如下的终端行为数据,

[0075] sip:192.168.130.241dip:192.168.131.125,dport:3306,trhour:10,resLen:1026,reqLen:10,interface:0001,action:0,duration:10

[0076] 其生成的向量为:

[0077] (1,1,3306,0,0001,10,12,4)。

[0078] (3)基于spark的Kmeans聚类分析模型的建立

[0079] 如图3,历史行为向量集如图3a所示。步骤a:选择聚类的个数为4,并对训练数据集中的数据进行归一化和向量化处理;步骤b:随机选取训练向量集中的4个向量作为聚类中心,随机选取的聚类中心如图3b中“+”所示;步骤c:计算训练数据集中每个向量与4个聚类中心的距离,并将该向量分配给与其距离最近的聚类中心,从而获得4个簇群;步骤d:对每个簇群计算该簇群所有数据点的平均值,并将其作为该簇群的新聚类中心;步骤e:重复c-d,直到聚类中心不再改变,稳定在如图3c的4个点,由此获得4个簇群。

[0080] 判断Kmeans聚类分析模型是否存在,如果不存在,以输入的终端历史行为数据(向量)作为输入数据,通过大量数据对模型进行训练,得到收敛的聚类结果,建立Kmeans聚类分析模型。

[0081] 将训练好的Kmeans聚类分析模型保存下来,对新采集到的终端行为数据计算其与各聚类分析中心的距离,设定一个阀值,当距离超过该阀值时,判断为该sip和dip为可疑行为数据。

[0082] (4)异常行为分析结果的攻击链地图展示

[0083] 对第(3)步中发现的可疑数据进行分析,找到该可疑sip、dip地址的ip地址链,并将ip地址与对应的位置信息关联上,并将该攻击链展示在地图上。

[0084] 实施例2

[0085] 如图2,本发明公开了一种网络入侵检测系统,包括:

[0086] 数据处理单元，用于对终端行为数据进行数据处理；

[0087] 判断单元，判断是否存在Kmeans聚类分析模型；

[0088] 聚类分析模型建立单元，用于根据历史终端行为数据建立基于spark的Kmeans聚类分析模型；

[0089] 数据分析单元，用于对新接收到的终端行为数据进行分析；

[0090] 异常行为分析结果展示单元，用于检测到网络入侵和可疑连接行为时，进行攻击链地图展示。

[0091] 其中数据处理单元用于通过数据采集模块对访问服务器的终端行为数据进行采集，并对采集的数据进行归一化处理和向量化处理，将每条终端行为记录转化为一个与请求相关的终端行为向量。

[0092] 其中聚类分析模型建立单元用于以历史终端行为向量作为输入数据，通过大量训练数据对聚类分析模型进行训练，得到收敛的聚类结果，保存训练好的聚类分析模型。

[0093] 数据分析单元用于对终端行为数据进行分析，具体为对新采集到的终端行为数据向量，计算其与各聚类分析中心的距离，设定一个阀值，当距离超过该阀值时，判断为该终端行为数据为可疑行为数据。

[0094] 异常行为分析结果展示单元对数据分析单元中发现的可疑数据进行分析，找到可疑行为数据对应的源ip地址和目的ip地址，将该源ip地址和目的ip地址与对应的位置信息关联，以攻击链的形式展示在地图上。

[0095] 某金融机构的大数据分析系统中对基于终端行为流量数据的威胁检测进行分析中进行应用，有效的对异常流量进行了报警、并对攻击链基于地理位置信息进行展示。

[0096] 对于本领域技术人员而言，显然本发明实施例不限于上述示范性实施例的细节，而且在不背离本发明实施例的精神或基本特征的情况下，能够以其他的具体形式实现本发明实施例。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明实施例的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明实施例内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统、装置或终端权利要求中陈述的多个单元、模块或装置也可以由同一个单元、模块或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

[0097] 最后应说明的是，以上实施方式仅用以说明本发明实施例的技术方案而非限制，尽管参照以上较佳实施方式对本发明实施例进行了详细说明，本领域的普通技术人员应当理解，可以对本发明实施例的技术方案进行修改或等同替换都不应脱离本发明实施例的技术方案的精神和范围。

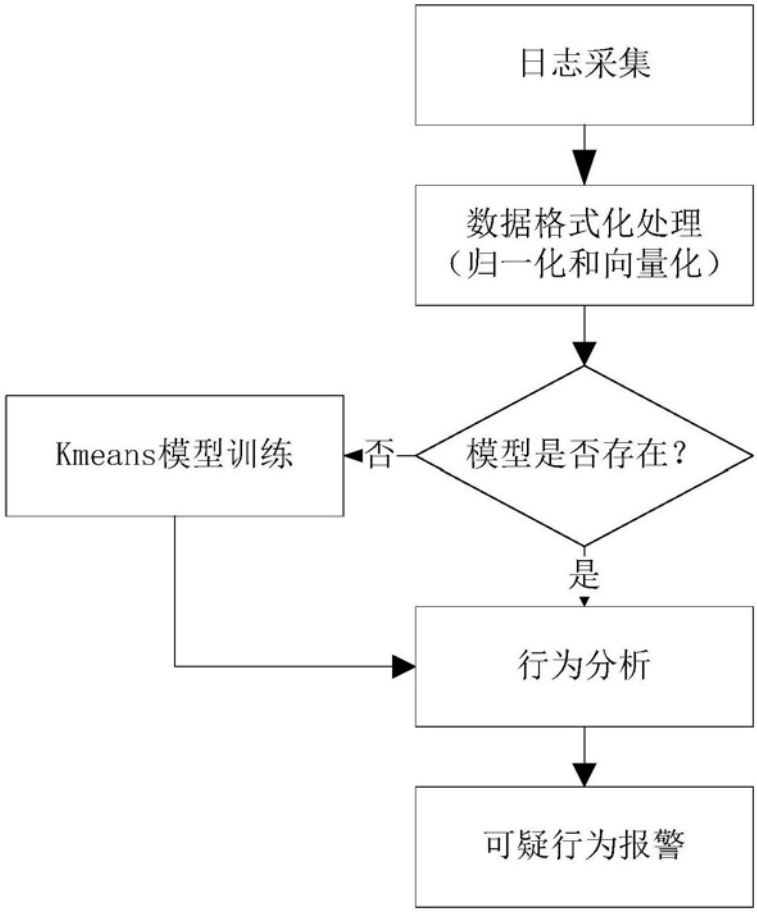


图1

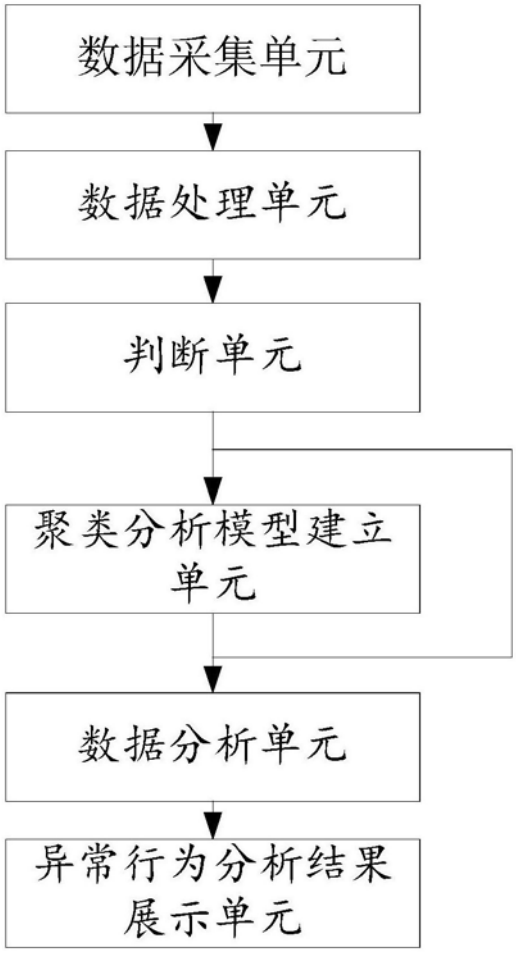


图2

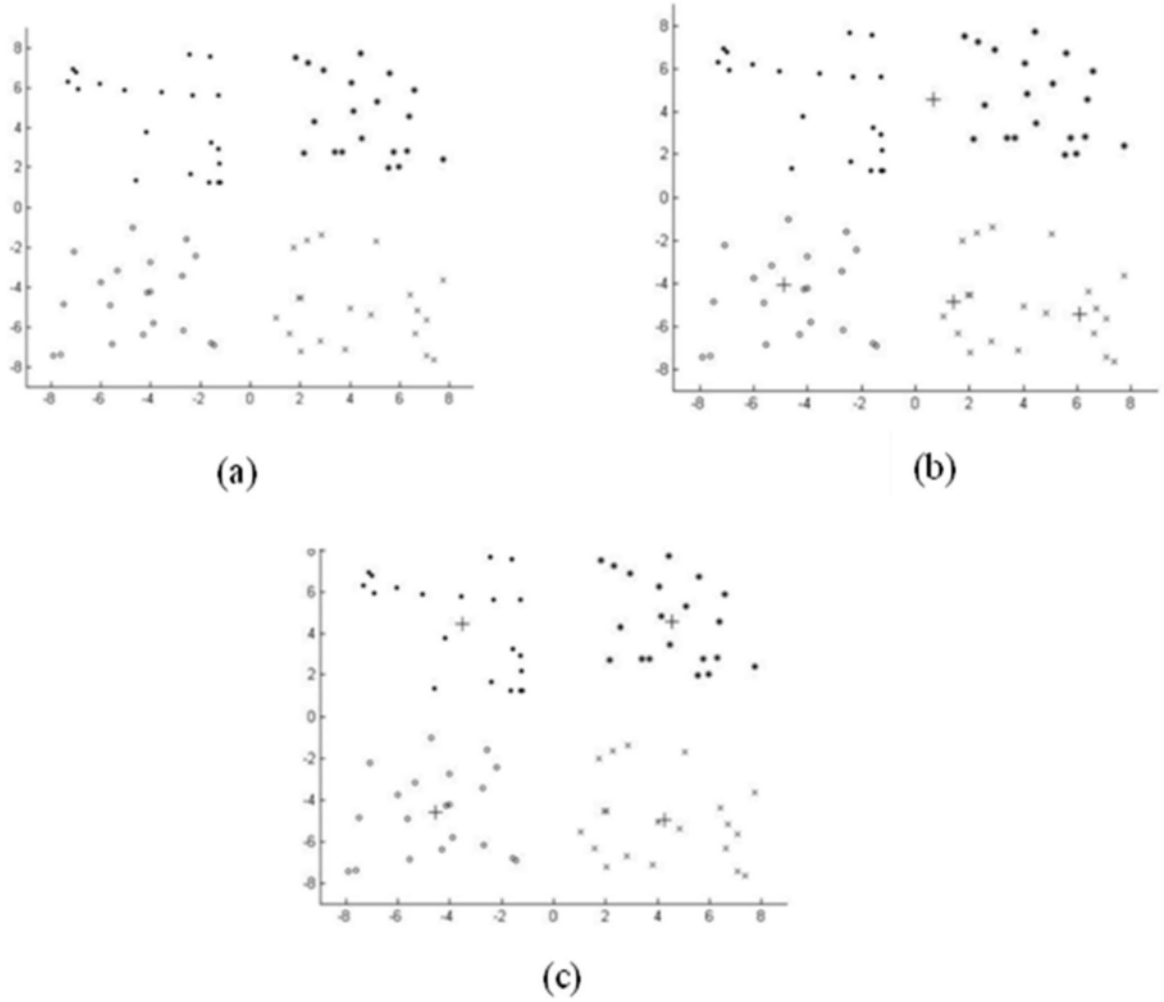


图3