(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104517052 B
(45)授权公告日 2017.05.10

(21)申请号 201410747764.7
(22)申请日 2014.12.09
(65)同一申请的已公布的文献号
申请公布号 CN 104517052 A
(43)申请公布日 2015.04.15
(73)专利权人 中国科学院深圳先进技术研究院
地址 518000 广东省深圳市南山区西丽大学城学苑大道1068号
(72)发明人 张爽 张涌 宁立
(74)专利代理机构 深圳中一专利商标事务所 44237
代理人 张全文
(51)Int.Cl.
*G06F 21/55*(2013.01)
*G06F 17/30*(2006.01)

审查员 张莹

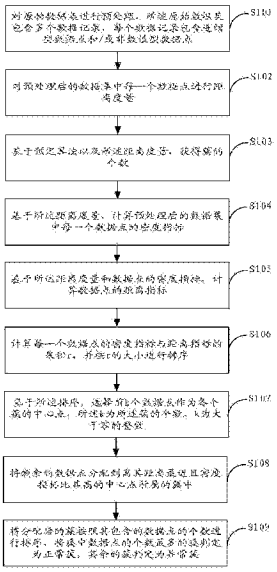
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种入侵检测方法及装置

(57)摘要

本发明适用于信息安全技术领域，提供了一种入侵检测方法及装置，所述方法包括：对原始数据集进行预处理；对预处理后的数据点进行距离度量；基于预定算法以及距离度量，获得簇的个数；基于距离度量，计算预处理后的数据点的密度指标；基于距离度量和密度指标，计算数据点的距离指标；计算数据点的密度指标与距离指标的乘积r并排序；选择前k个数据点作为各个簇的中心点；将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。通过本发明，可有效解决现有技术存在的运算开销大，初始值的设定影响聚类结果的问题。




CN 104517052 B

1.一种入侵检测方法,其特征在于,所述方法包括:

对原始数据集进行预处理,所述原始数据集包含多个数据记录,每个数据记录包含连续型数据点和/或非数值型数据点;

对预处理后的数据集中每一个数据点进行距离度量;

基于预定算法以及所述距离度量,获得簇的个数;

基于所述距离度量,计算预处理后的数据集中每一个数据点的密度指标;

基于所述距离度量和数据点的密度指标,计算数据点的距离指标;

计算每一个数据点的密度指标与距离指标的乘积r,并按r值从大到小进行排序;

基于所述排序,选择前k个数据点作为各个簇的中心点,所述k为所述簇的个数,k为大于零的整数;

将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中;

将分配后的簇按照其包含的数据点的个数进行排序,将簇中数据点的个数最多的簇判定为正常簇,其余的簇判定为异常簇;

所述基于所述距离度量,计算预处理后的数据集中每一个数据点的密度指标包括:

针对某个数据点i,计算i与其周围数据点的距离,将距离小于或等于预定距离的周围数据点的个数作为所述i的密度指标;

所述基于所述距离度量和数据点的密度指标,计算数据点的距离指标包括:

针对某个数据点i,获取密度指标比i密度指标大的数据点$M_j$,并计算i与$M_j$的距离,将计算得到的最小距离作为所述i的距离指标,其中j大于或等于1。

2.如权利要求1所述的方法,其特征在于,所述对原始数据集进行预处理包括:

对原始数据集中的连续型数据点,将其数据取值从[min,max]映射到范围小于预设值的区间;

对原始数据集中的非数值型数据点,将其离散化后,通过编码映射成数值,或者直接在所述距离度量中进行比较。

3.如权利要求1或2所述的方法,其特征在于,所述对预处理后的数据集中每一个数据点进行距离度量包括:

基于加权的欧几里德公式对预处理后的数据集中每一个数据点进行距离度量。

4.一种入侵检测装置,其特征在于,所述装置包括:

预处理单元,用于对原始数据集进行预处理,所述原始数据集包含多个数据记录,每个数据记录包含连续型数据点和/或非数值型数据点;

距离度量单元,用于对预处理后的数据集中每一个数据点进行距离度量;

簇个数获取单元,用于基于预定算法以及所述距离度量,获得簇的个数;

密度指标计算单元,用于基于所述距离度量,计算预处理后的数据集中每一个数据点的密度指标;

距离指标计算单元,用于基于所述距离度量和数据点的密度指标,计算数据点的距离指标;

排序单元,用于计算每一个数据点的密度指标与距离指标的乘积r,并按r值从大到小进行排序;

中心点确定单元,用于基于所述排序,选择前k个数据点作为各个簇的中心点,所述k为

所述簇的个数，k为大于零的整数；

分配单元，用于将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；

判定单元，用于将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇；

所述密度指标计算单元具体用于：

针对某个数据点i，计算i与其周围数据点的距离，将距离小于或等于预定距离的周围数据点的个数作为所述i的密度指标；

所述距离指标计算单元具体用于：

针对某个数据点i，获取密度指标比i密度指标大的数据点$M_j$，并计算i与$M_j$的距离，将计算得到的最小距离作为所述i的距离指标，其中j大于或等于1。

5.如权利要求4所述的装置，其特征在于，所述预处理单元具体用于：

对原始数据集中的连续型数据点，将其数据取值从[min，max]映射到范围小于预设值的区间；

对原始数据集中的非数值型数据点，将其离散化后，通过编码映射成数值，或者直接在所述距离度量中进行比较。

6.如权利要求4或5所述的装置，其特征在于，所述距离度量单元具体用于：

基于加权的欧几里德公式对预处理后的数据集中每一个数据点进行距离度量。

# 一种入侵检测方法及装置

## 技术领域

[0001] 本发明属于信息安全技术领域，尤其涉及一种入侵检测方法及装置。

## 背景技术

[0002] 现有应用到入侵检测中的聚类算法大致分为两种：一种是基于划分的的聚类算法，一种是基于密度的的聚类算法。

[0003] 基于划分的聚类算法，如K-means，由于簇的个数K与初始聚类中心点是事先人为选定的，一旦选择不好，可能无法获得有效的聚类结果；其次，基于划分的聚类算法不能处理非球形簇、不同尺寸和不同密度的簇。

[0004] 基于密度的聚类算法，如经典的DBSCAN(Density-Based Spatial Clustering of Applications with Noise)，对于高维度且数据量较大的入侵数据，运算开销会比较大，而且预先定义的密度阈值会对后面的聚类结果有明显的影响。

## 发明内容

[0005] 鉴于此，本发明实施例提供一种入侵检测方法及装置，以解决现有技术存在的运算开销大，初始值的设定影响聚类结果的问题。

[0006] 一方面，本发明实施例提供一种入侵检测方法，所述方法包括：

[0007] 对原始数据集进行预处理，所述原始数据集包含多个数据记录，每个数据记录包含连续型数据点和/或非数值型数据点；

[0008] 对预处理后的数据集中每一个数据点进行距离度量；

[0009] 基于预定算法以及所述距离度量，获得簇的个数；

[0010] 基于所述距离度量，计算预处理后的数据集中每一个数据点的密度指标；

[0011] 基于所述距离度量和数据点的密度指标，计算数据点的距离指标；

[0012] 计算每一个数据点的密度指标与距离指标的乘积r，并按r的大小进行排序；

[0013] 基于所述排序，选择前k个数据点作为各个簇的中心点，所述k为所述簇的个数，k为大于零的整数；

[0014] 将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；

[0015] 将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。

[0016] 另一方面，本发明实施例提供一种入侵检测装置，所述装置包括：

[0017] 预处理单元，用于对原始数据集进行预处理，所述原始数据集包含多个数据记录，每个数据记录包含连续型数据点和/或非数值型数据点；

[0018] 距离度量单元，用于对预处理后的数据集中每一个数据点进行距离度量；

[0019] 簇个数获取单元，用于基于预定算法以及所述距离度量，获得簇的个数；

[0020] 密度指标计算单元，用于基于所述距离度量，计算预处理后的数据集中每一个数据点的密度指标；

[0021] 距离指标计算单元，用于基于所述距离度量和数据点的密度指标，计算数据点的距离指标；

[0022] 排序单元，用于计算每一个数据点的密度指标与距离指标的乘积r，并按r的大小进行排序；

[0023] 中心点确定单元，用于基于所述排序，选择前k个数据点作为各个簇的中心点，所述k为所述簇的个数，k为大于零的整数；

[0024] 分配单元，用于将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；

[0025] 判定单元，用于将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。

[0026] 本发明实施例与现有技术相比存在的有益效果是：本发明实施例基于预定算法(例如Canopy算法)以及距离度量(例如加权的欧几里德距离度量)，获得簇的个数，并通过计算获得预处理后的数据集中每一个数据点的密度指标和距离指标，将所述密度指标和距离指标的乘积作为综合指标，根据所述综合指标获得簇的中心点，解决了现有技术人为设定初始值(如簇的中心点、簇的个数等)影响聚类结果的问题。而且，对于高维度且数据量较大的入侵数据，相比于现有的聚类方法，无需迭代最优目标函数，明显减少了计算开销。另外，由于是基于密度的聚类算法，对于非球形簇，也有很好的聚类效果，并能自动检测出异常簇，具有较强的易用性和实用性。

## 附图说明

[0027] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0028] 图1是本发明实施例一提供的入侵检测方法的实现流程图；

[0029] 图2是本发明实施例二提供的入侵检测装置的组成结构图。

## 具体实施方式

[0030] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0031] 为了说明本发明所述的技术方案，下面通过具体实施例来进行说明。

[0032] 实施例一：

[0033] 图1示出了本发明实施例一提供的入侵检测方法的实现流程，该方法过程详述如下：

[0034] 在步骤S101中，对原始数据集进行预处理，所述原始数据集包含多个数据点。

[0035] 在本发明实施例中，所述原始数据集包含多个数据记录(例如异构型数据记录)，每个数据记录可能包含连续型数据点和/或非数值型数据点，需要对二者分别进行数据规范化处理，具体可以是：

[0036] 对原始数据集中的连续型数据点，将其数据取值从[min，max]映射到范围小于预设值的区间(例如[0，1]区间)；

[0037] 对原始数据集中的非数值型数据，将其离散化后，通过编码映射成数值，或者直接在所述距离度量中进行比较，根据特定公式计算其距离。

[0038] 所述原始数据集经过上述数据规范化预处理后变成高维向量组。其中，所述原始数据集可以为KDD CUP99数据集，该数据集分为训练数据集与检测数据集，其中包含了大量的数据记录，每个数据记录含有41维特征，共有39种类型的攻击记录，训练数据集中每个数据记录都被标记为正常或某种攻击，其中有22种攻击类型的记录。另有17种未知攻击类型出现在测试数据集中。

[0039] 需要说明的是，本发明实施例对数据集进行规范化处理，将属性数据按比例缩放，使之落入一个小的特定区间，对于涉及距离度量的聚类算法，将有助于加快学习阶段的速度，并且可以帮助防止具有较大初始值域的属性与具有较小初始值域的属性相比权重过大，进而影响距离度量的准确性。

[0040] 在步骤S102中，对预处理后的数据集中每一个数据点进行距离度量。

[0041] 由于在密度聚类算法中，数据量较大、特征维数较多的数据在运算方面一般开销较大。因此，本发明实施例基于欧几里德公式对预处理后的数据进行距离度量，采用欧几里德公式的突出优点是计算简单，运行速度快，且可以支持多维空间索引，欧几里德公式具体如下：

[0042] $$d(i,j)=\sqrt{\left|x_{i1}-x_{j1}\right|^2+\left|x_{i2}-x_{j2}\right|^2+...+\left|x_{ip}-x_{jp}\right|^2}$$ 。

[0043] 另外，为了体现不同属性的权重，本实施例可以给不同的属性赋以不同的权值，即基于加权的欧几里德公式对预处理后的数据进行距离度量，公式具体如下：

[0044] $$d(i,j)=\sqrt{w_1\left|x_{i1}-x_{j1}\right|^2+w_2\left|x_{i2}-x_{j2}\right|^2+...+w_p\left|x_{ip}-x_{jp}\right|^2}$$

[0045] 其中，$(x_{i1},x_{i2},...,x_{ip})$为数据$x_i$的属性向量，$(x_{j1},x_{j2},...,x_{jp})$为数据$x_j$的属性向量，$w_p$为对应的权值，p为大于0的整数。

[0046] 在步骤S103中，基于预定算法以及所述距离度量，获得簇的个数。

[0047] 在本发明实施例中，所述预定算法包括但不限于Canopy算法。本发明实施例以Canopy算法为例进行说明：将经过预处理后的原始数据集作为一个集合A，设置一个值T，T的取值为两两数据点之间距离的平均数；从集合A中任意选择一个数据点作为基点X，根据距离度量公式，计算数据集中其他数据点与所述基点X之间的距离；若某个数据点与所述基点X的距离小于T，则将此数据点与所述基点X划为一个Canopy(即相似数据点的集合)，该Canopy最终将变为与所述基点X距离小于T的数据点的一个子集合；将子集合中的数据点剔除集合A，继续选择另一个基点Y，计算集合A中剩余数据点与基点Y的距离，从而获得第二个Canopy；重复上述步骤，将最终获得的Canopy的个数作为聚类后得到的簇的个数，簇的个数即为K(K为大于零的整数)。K值将会作为后续的改进聚类算法中的一个输入参数。

[0048] 本发明实施例基于Canopy算法以及所述距离度量，自动获得簇的个数，解决了现有技术人为设定簇的个数影响聚类结果的问题。而且采用Canopy算法可有效提高聚类的速度。

[0049] 在步骤S104中，基于所述距离度量，计算预处理后的数据集中每一个数据点的密度指标。

[0050] 具体的可以是，针对某个数据点i，计算i与其周围数据点(预设范围内的数据点)的距离，将距离小于或等于预定距离的周围数据点的个数作为所述i的密度指标。

[0051] 在步骤S105中，基于所述距离度量和数据点的密度指标，计算数据点的距离指标。

[0052] 具体的可以是，针对某个数据点i，获取密度指标比i密度指标大的数据点$M_j$，并计算i与$M_j$的距离，将计算得到的最小距离作为所述i的距离指标，其中j大于或等于1。

[0053] 在步骤S106中，计算每一个数据点的密度指标与距离指标的乘积r，并按r的大小进行排序。

[0054] 在本发明实施例中，所述乘积r可作为数据点的综合指标。其中，r值越大，说明该数据点为簇的中心点的可能性越大。

[0055] 在步骤S107中，基于所述排序，选择前k个数据点作为各个簇的中心点，所述k为所述簇的个数。

[0056] 在本发明实施例中，簇的个数为k个，簇的中心点也为k个，每个簇对应一个中心点。其中，k为大于零的整数。

[0057] 在步骤S108中，将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；

[0058] 在步骤S109中，将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。

[0059] 聚类的目的是要将一个数据集划分为若干组，使得组内的相似性大于组间相似性。本发明实施例在入侵检测过程中，采用改进的密度聚类算法进行分析，即在经过计算密度指标、距离指标、综合指标、获得簇的中心点、将剩余点进行分配几个步骤后，便可以将相似的数据点划分到同一组内。另外，由于入侵数据集中正常行为的簇所包含的数据点在数量上远远大于非正常行为的簇所包含的数据点，因此本实施例将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。可选的，还可以预先设定一数值，将簇中数据点的个数大于或等于所述预定数值的簇判定为正常簇，小于所述预定数值的簇判定为异常簇。

[0060] 进一步的，本发明实施例还包括：

[0061] 根据聚类后的结果进行聚类评测，评测指标为正确率与误检率。其中，正确率表示检测到异常数据点的个数与原始数据集中异常数据点的总数的比值；误检率表示被误认为异常数据点的个数与原始数据集中正常数据点的总数的比值。本发明实施例根据所述测评指标，对k值进行自适应调整，并在调整后，再次进行聚类运算，得到新的聚类结果与新的评测标准，直到得到最佳的聚类结果(即正确率最高，误检率最低)。可选的，还可以预先设定第一阈值以及第二阈值，在所述正确率大于所述第一阈值，所述误检率小于第二阈值时，输出聚类结果。

[0062] 本发明实施例改进的密度聚类算法基于的条件是：1)一个簇是由中心点与边界点组合而成，而且中心点的密度值比边界点的密度值要大；2)簇与簇之间都有一定的距离。本发明实施例基于所述条件，通过计算数据点的综合指标(即密度指标与距离指标的乘积)，得到簇的中心点，即密度较高、相对距离较大的点。然后将剩余的点按距离分配到比本身密

度高的中心点所属的簇中，得到簇的中心点与簇的类数等相关信息，进而划分出正常数据与异常数据。而且可以对对k值进行自适应调整，从而获得更佳的聚类结果。

[0063] 实施例二：

[0064] 图2示出了本发明实施例二提供的入侵检测装置的组成结构，为了便于说明，仅示出了与本发明实施例相关的部分。

[0065] 该入侵检测装置可以是运行于各终端设备(例如手机、平板电脑等)内的软件单元、硬件单元或者软硬件相结合的单元，也可以作为独立的挂件集成到所述终端设备中或者运行于所述终端设备的应用系统中。

[0066] 该入侵检测装置包括：

[0067] 预处理单元21，用于对原始数据集进行预处理，所述原始数据集包含多个数据记录，每个数据记录包含连续型数据点和/或非数值型数据点；

[0068] 距离度量单元22，用于对预处理后的数据集中每一个数据点进行距离度量；

[0069] 簇个数获取单元23，用于基于预定算法以及所述距离度量，获得簇的个数；

[0070] 密度指标计算单元24，用于基于所述距离度量，计算预处理后的数据集中每一个数据点的密度指标；

[0071] 距离指标计算单元25，用于基于所述距离度量和数据点的密度指标，计算数据点的距离指标；

[0072] 排序单元26，用于计算每一个数据点的密度指标与距离指标的乘积r，并按r的大小进行排序；

[0073] 中心点确定单元27，用于基于所述排序，选择前k个数据点作为各个簇的中心点，所述k为所述簇的个数，k为大于零的整数；

[0074] 分配单元28，用于将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中；

[0075] 判定单元29，用于将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇。

[0076] 进一步的，所述预处理单元21具体用于：

[0077] 对原始数据集中的连续型数据点，将其数据取值从[min,max]映射到范围小于预设值的区间；

[0078] 对原始数据集中的非数值型数据点，将其离散化后，通过编码映射成数值，或者直接在所述距离度量中进行比较。

[0079] 进一步的，所述距离度量单元22具体用于：

[0080] 基于加权的欧几里德公式对预处理后的数据集中每一个数据点进行距离度量。

[0081] 进一步的，所述密度指标计算单元24具体用于：

[0082] 针对某个数据点i，计算i与其周围数据点的距离，将距离小于或等于预定距离的周围数据点的个数作为所述i的密度指标。

[0083] 进一步的，所述距离指标计算单元25具体用于：

[0084] 针对某个数据点i，获取密度指标比i密度指标大的数据点$M_j$，并计算i与$M_j$的距离，将计算得到的最小距离作为所述i的距离指标，其中j大于或等于1。

[0085] 所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功

能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中，上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。另外，各功能单元、模块的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。上述装置中单元、模块的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0086] 综上所述，本发明实施例通过1)对数据集进行规范化处理，将属性数据按比例缩放，使之落入一个小的特定区间，对于涉及距离度量的聚类算法，将有助于加快学习阶段的速度，并且可以帮助防止具有较大初始值域的属性与具有较小初始值域的属性相比权重过大，进而影响距离度量的准确性；2)基于Canopy算法以及加权的欧几里德距离度量，获得簇的个数，并通过计算获得预处理后的数据集中每一个数据点的密度指标和距离指标，将所述密度指标和距离指标的乘积作为综合指标，根据所述综合指标获得簇的中心点，解决了现有技术人为设定初始值(如簇的中心点、簇的个数等)影响聚类结果的问题；3)对于高维度且数据量较大的入侵数据，相比于现有的聚类方法，无需迭代最优目标函数，明显减少了计算开销。另外，由于是基于密度的聚类算法，对于非球形簇，也有很好的聚类效果，并能自动检测出异常簇，具有较强的易用性和实用性，具有较强的易用性和实用性。

[0087] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0088] 在本发明所提供的实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。

[0089] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0090] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0091] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台

计算机设备(可以是个人计算机,服务器,或者网络设备等)或处理器(processor)执行本发明实施例各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0092] 以上所述实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明实施例各实施例技术方案的精神和范围。

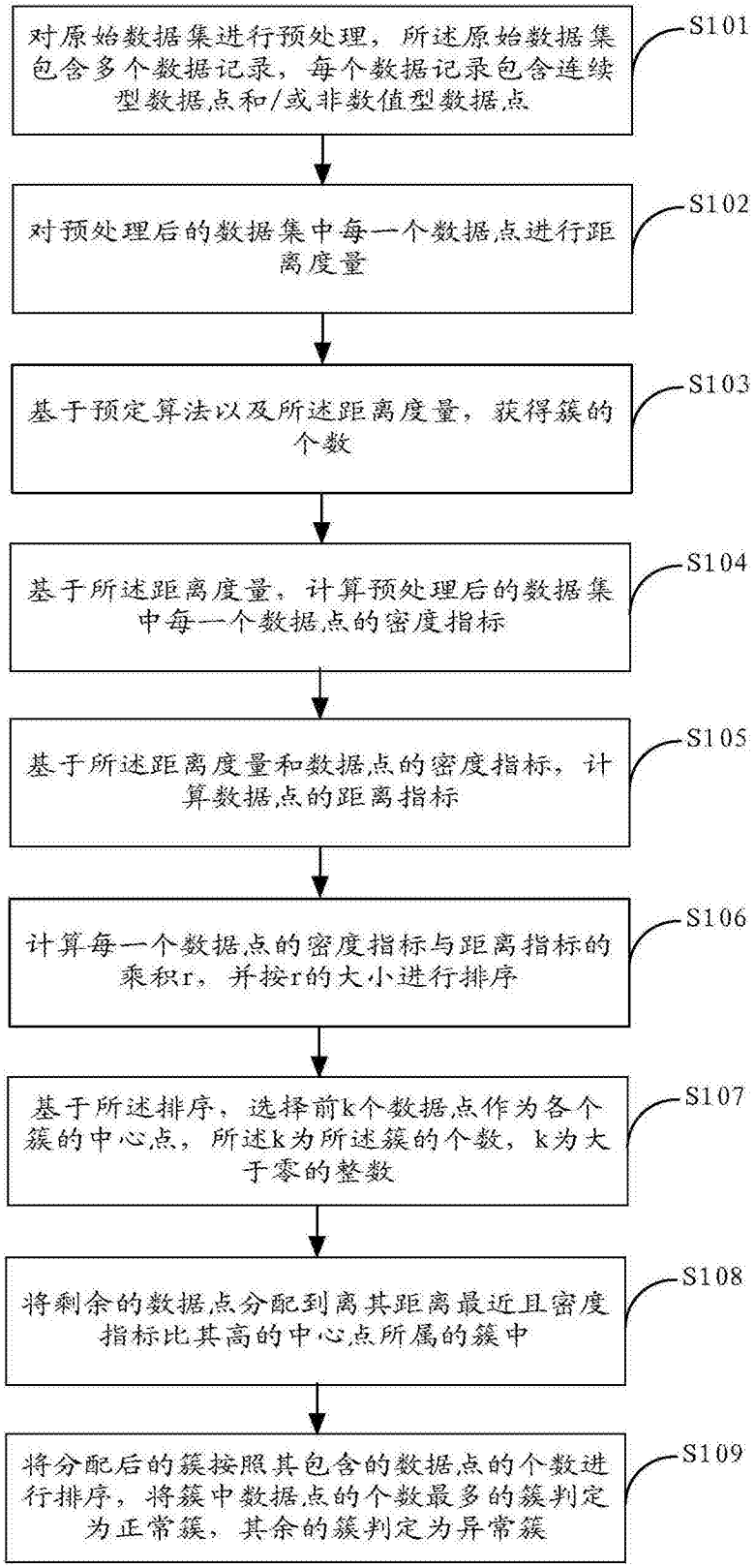

对原始数据集进行预处理，所述原始数据集包含多个数据记录，每个数据记录包含连续型数据点和/或非数值型数据点
S101
对预处理后的数据集中每一个数据点进行距离度量
S102
基于预定算法以及所述距离度量，获得簇的个数
S103
基于所述距离度量，计算预处理后的数据集中每一个数据点的密度指标
S104
基于所述距离度量和数据点的密度指标，计算数据点的距离指标
S105
计算每一个数据点的密度指标与距离指标的乘积r，并按r的大小进行排序
S106
基于所述排序，选择前k个数据点作为各个簇的中心点，所述k为所述簇的个数，k为大于零的整数
S107
将剩余的数据点分配到离其距离最近且密度指标比其高的中心点所属的簇中
S108
将分配后的簇按照其包含的数据点的个数进行排序，将簇中数据点的个数最多的簇判定为正常簇，其余的簇判定为异常簇
S109

图1

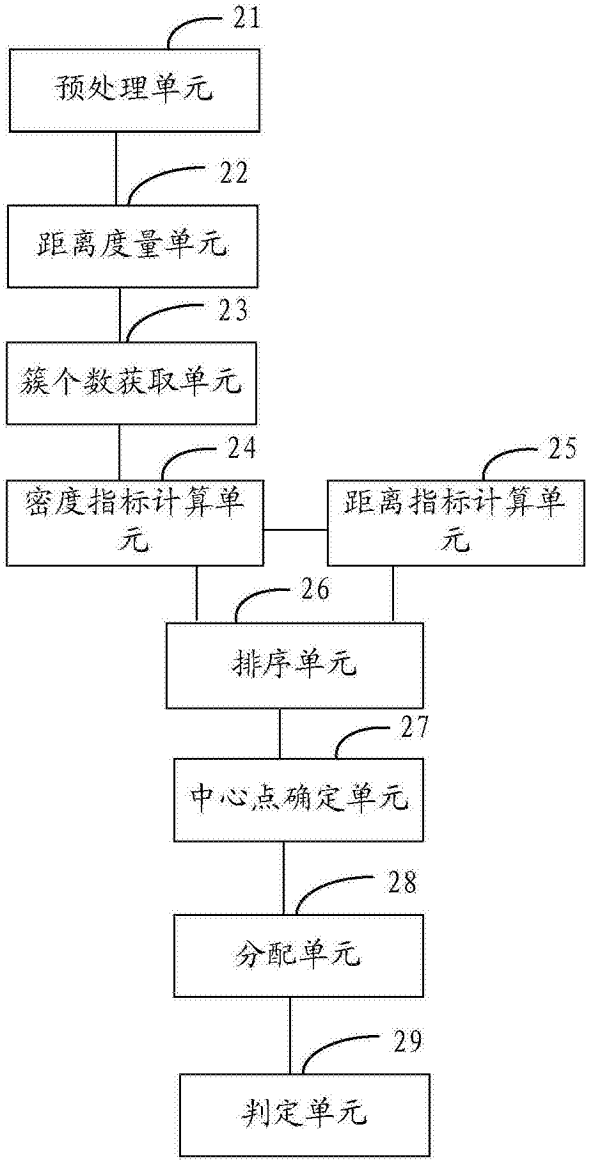

21
预处理单元
22
距离度量单元
23
簇个数获取单元
24
密度指标计算单元
25
距离指标计算单元
26
排序单元
27
中心点确定单元
28
分配单元
29
判定单元

图2