

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 8 月 6 日 (06.08.2020)



(10) 国际公布号
WO 2020/155755 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2019/117355

(22) 国际申请日: 2019 年 11 月 12 日 (12.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910079172.5 2019 年 1 月 28 日 (28.01.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 杨志鸿(YANG, Zhihong); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 阮晓雯(RUAN, Xiaowen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路 6009 号绿景广场 B 栋 20 层 B, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: SPECTRAL CLUSTERING-BASED OPTIMIZATION METHOD FOR ANOMALY POINT RATIO, DEVICE, AND COMPUTER APPARATUS

(54) 发明名称: 基于谱聚类的异常点比例优化方法、装置及计算机设备

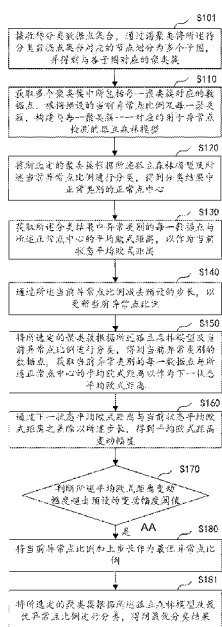


图 1

S101 Receive a data point set to be classified, and divide nodes corresponding to the data point set into multiple sub-graphs by means of spectral clustering, and obtain respective clusters corresponding to the sub-graphs

S110 Acquire corresponding data points of each of the multiple clusters, and construct, according to a pre-determined current anomaly point ratio and the clusters, isolation forest models in one-to-one correspondence with the clusters, the isolation forest models being used for anomaly point detection

S120 Classify a selected cluster according to the isolation forest models and the current anomaly point ratio, and obtain a normal point center of a normal type in a classification result

S130 Acquire an average Euclidean distance of each data point being an anomaly type in the classification result to the normal point center, and use the same as an average Euclidean distance of a current state

S140 Subtract a pre-determined step size from the current anomaly point ratio to update the current anomaly point ratio

S150 Use the current anomaly point ratio plus the step size as an optimal anomaly point ratio

S181 Classify a selected cluster according to the isolation forest models and the optimal anomaly point ratio, and obtain an optimal classification result

AA Yes

(57) Abstract: The present application discloses a spectral clustering-based optimization method for an anomaly point ratio, a device, and a computer apparatus. The method comprises: receiving a data point set to be classified, and performing clustering on the data point set by means of spectral clustering to obtain multiple clusters; acquiring data points corresponding to each of the clusters, and constructing, according to a pre-determined current anomaly point ratio and the clusters, isolation forest models in one-to-one correspondence with the clusters; continuously adjusting the current anomaly point ratio until a change extent of an average Euclidean distance exceeds

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

a change extent threshold, and using the current anomaly point ratio plus a step size as an optimal anomaly point ratio; and classifying a selected cluster according to the optimal anomaly point ratio, and obtaining an optimal classification result.

(57) 摘要: 本申请公开了基于谱聚类的异常点比例优化方法、装置及计算机设备。该方法包括: 通过接收待分类数据点集合, 通过谱聚类将待分类数据点集合进行聚类得到多个聚类簇; 获取各聚类簇对应的数据点, 根据预设的当前异常点比例及各聚类簇, 构建与各聚类簇一一对应的孤立森林模型; 通过不断调整当前异常点比例, 直至平均欧式距离变动幅度超出变动幅度阈值, 将当前异常点比例加上步长作为最优异常点比例; 将所选定的聚类簇根据最优异常点比例进行分类, 得到最优分类结果。

发明名称：基于谱聚类的异常点比例优化方法、装置及计算机设备

本申请要求于2019年1月28日提交中国专利局、申请号为201910079172.5、申请名称为“基于谱聚类的异常点比例优化方法、装置及计算机设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及智能决策技术领域，尤其涉及一种基于谱聚类的异常点比例优化方法、装置及计算机设备。

背景技术

异常值分析是检验数据是否有录入错误以及含有不合常理的数据的过程，忽视异常值的存在是十分危险的，不加剔除地把异常值包括进数据的计算分析过程中，对结果会产生不良影响。目前，当云服务器接收了海量的数据并需对其进行异常点检测是，若仅将海量数据视为只有一个中心的数据集，会导致用于异常点检测的无监督模型的区分效果较差，无法准确的检测出异常点。而且无监督模型在检测前的异常点比例的设置依赖用户经验，导致设置难度较大。

发明内容

本申请实施例提供了一种基于谱聚类的异常点比例优化方法、装置及计算机设备，旨在解决现有技术中海量数据视为只有一个中心的数据集，会导致用于异常点检测的无监督模型的区分效果较差，无法准确的检测出异常点的问题。

第一方面，本申请实施例提供了一种基于谱聚类的异常点比例优化方法，其包括：

接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇；

获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型；

将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，

得到分类结果中正常类别的正常点中心；

获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离；

通过所述当前异常点比例减去预设的步长，以更新当前异常点比例；

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离；

通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度；

判断所述平均欧式距离变动幅度超出预设的变动幅度阈值；

若所述平均欧式距离变动幅度超出所述变动幅度阈值，将当前异常点比例加上步长作为最优异常点比例；以及

将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果。

第二方面，本申请实施例提供了一种基于谱聚类的异常点比例优化装置，其包括：

谱聚类单元，用于接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇；

孤立森林模型训练单元，用于获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型；

正常点中心获取单元，用于将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心；

第一平均距离计算单元，用于获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离；

第一异常点比例更新单元，用于通过所述当前异常点比例减去预设的步长，以更新当前异常点比例；

第二平均距离计算单元，用于将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离；

平均距离变动幅度获取单元，用于通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度；

幅度判断单元，用于判断所述平均欧式距离变动幅度超出预设的变动幅度阈值；

最优比例获取单元，用于若所述平均欧式距离变动幅度超出所述变动幅度阈值，将当前异常点比例加上步长作为最优异常点比例；以及

最优分类单元，用于将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果。

第三方面，本申请实施例又提供了一种计算机设备，其包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述第一方面所述的基于谱聚类的异常点比例优化方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，其中所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行时使所述处理器执行上述第一方面所述的基于谱聚类的异常点比例优化方法。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请实施例提供的基于谱聚类的异常点比例优化方法的流程示意图；

图 2 为本申请实施例提供的基于谱聚类的异常点比例优化方法的子流程示意图；

图 3 为本申请实施例提供的基于谱聚类的异常点比例优化方法的另一子流程示意图；

图 4 为本申请实施例提供的基于谱聚类的异常点比例优化方法的另一子流程示意图；

图 5 为本申请实施例提供的基于谱聚类的异常点比例优化方法的另一流程

示意图；

图 6 为本申请实施例提供的基于谱聚类的异常点比例优化装置的示意性框图；

图 7 为本申请实施例提供的基于谱聚类的异常点比例优化装置的子单元示意性框图；

图 8 为本申请实施例提供的基于谱聚类的异常点比例优化装置的另一子单元示意性框图；

图 9 为本申请实施例提供的基于谱聚类的异常点比例优化装置的另一子单元示意性框图；

图 10 为本申请实施例提供的基于谱聚类的异常点比例优化装置的另一示意性框图；

图 11 为本申请实施例提供的计算机设备的示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

还应当理解，在此本申请说明书中所使用的术语仅仅是出于描述特定实施例的目的而并不意在限制本申请。如在本申请说明书和所附权利要求书中所使用的那样，除非上下文清楚地指明其它情况，否则单数形式的“一”、“一个”及“该”意在包括复数形式。

还应当进一步理解，在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

请参阅图 1，图 1 为本申请实施例提供的基于谱聚类的异常点比例优化方法的流程示意图，该基于谱聚类的异常点比例优化方法应用于服务器中，该方法

通过安装于服务器中的应用软件进行执行。

如图 1 所示，该方法包括步骤 S101~S181。

S101、接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇。

在本实施例中，当企业的云服务器接收了各用户上传的海量用户数据后，这些用户数据可视为待分类数据点集合。例如，待分类数据点集合可以是用户的保单数据，至少包括投保人姓名、投保人年龄、投保人保单数量、投保金额、投保年限、投保人手机号码等字段。此时可有选择性的选择其中一个字段数据作为主数据，而剩余的字段则作为上述主字段的属性数据。例如投保年限字段作为主数据，投保人的电话号码、身份证号等字段作为其属性数据。将待分类数据点集合中每一数据点转化为节点之后，可以通过谱聚类对节点进行划分，得到多个子图，每一个子图可以视为一个聚类簇。

在一实施例中，如图 2 所示，步骤 S101 包括：

S1011、获取所输入的相似度矩阵和目标聚类数目；

S1012、根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似矩阵；

S1013、根据所述相似矩阵构建邻接矩阵和对角矩阵，由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵；

S1014、获取所述拉普拉斯矩阵中的多个特征值的排名，若判断特征值的排名位于预设排名阈值之前，获取对应的特征向量以组成目标特征向量集；

S1015、将目标特征向量集合中每一特征向量转置为列向量并依次组合，以得到目标向量矩阵；

S1016、通过 k-means 算法将目标向量矩阵中各行向量进行聚类，得到与所述目标聚类数目相同的子图，并得到与各子图对应的聚类簇。

谱聚类是一种基于图论的聚类方法，通过对样本数据的拉普拉斯矩阵的特征向量进行聚类，从而达到对样本数据聚类的目的。谱聚类可以理解为将高维空间的数据映射到低维，然后在低维空间用其它聚类算法（如 k-means）进行聚类。

为了实现对高维空间的理赔数据映射到低维空间，需将所述理赔数据对应的节点先根据式（1）进行相似矩阵的构建：

$$s_{ij} = s(x_i, x_j) = \sum_{i=1, j=1}^n \exp \frac{-\|x_i - x_j\|}{2\sigma^2} \quad (\text{式 } 1)$$

其中, n 为赔数据对应的节点个数, x_i 和 x_j 分别表示任意一个节点, σ 表示节点的标准差, s_{ij} 则组成了相似矩阵。

由所输入的相似度矩阵来构建与所述理赔数据对应的节点相应的相似矩阵有 ϵ -邻近法, K 邻近法和全连接法。例如, 全连接法的计算公式如式 1。

之后根据式 2 来计算对角矩阵, 式 2 具体如下:

$$d_i = \sum_{j=1}^n w_{ij} \quad (\text{式 } 2)$$

其中, d_i 表示相似矩阵中每一行的元素之和, 由 d_i 组成对角矩阵 w_{ij} 则表示相似矩阵中第 i 行第 j 列的元素。

当由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵后, 即可以拉普拉斯矩阵中对应的每一特征向量转置为列向量, 从而组成目标向量矩阵。最后通过 k -means 算法将目标向量矩阵中各行向量进行聚类, 得到与所述目标聚类数目相同的子团。

S110、获取多个聚类簇中所包括每一聚类簇对应的数据点, 根据预设的当前异常点比例及每一聚类簇, 构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型。

在本实施例中, 例如, 云服务器接收了业务端所上传的待分类数据点集合并完成谱聚类分组后, 此时若所设置初始的当前异常点比例为 0.5 (如将初始的当前异常点比例记为 m_0), 表示所期望的孤立森林模型的分类结果中正常点样本和异常点样本比例为 1:1。由于假设正常点数量比异常点多, 因此此时异常点类别中含有大量的错分正常点。当异常点比例减少的时候, 异常点类别中的正常点会被剔除。此时, 先根据预设的当前异常点比例及每一聚类簇分别构建用于异常点检测的孤立森林模型, 作为后续调整当前异常点比例并重新分类的模型基础。

在一实施例中, 如图 3 所示, 步骤 S110 包括:

S111、从各聚类簇中均随机获取一个数据属性, 及由各聚类簇中所选定的数据属性和当前异常点比例所确定的分裂值;

S112、根据所述数据属性及所述分裂值将各聚类簇分别进行划分, 得到各聚类簇分别对应的多个孤立树, 以组合得到与各聚类簇对应的用于异常点检测

的孤立森林模型。

在本实施例中，例如聚类簇 1 对应的数据集为 $D1$ ，其中 $D1=\{d_1, d_2, \dots, d_n\}$ ，从中随机选择一个数据属性 A ，并由数据属性 A 和当前异常点比例确定一个分裂值 p_1 ；然后对训练数据集中每个数据对象 d_i ，按照数据属性 A 的分裂值 p_1 进行划分。若 $d_i(A)$ 小于 p_1 ，则放在左子树，反之则在右子树。此时再随机选择一个数据属性 B ，并由数据属性 B 和当前异常点比例确定一个分裂值 p_2 ；然后对左子树和右子树均根据按照数据属性 B 的分裂值 p_2 进行划分，得到与左子树对应的次级左子树和次级右子树，以及与右子树对应的次级左子树和次级右子树。以此迭代，直至满足一下条件之一：（1） $D1$ 中剩下一条数据或者多条相同的数据；（2）孤立树达到最大高度。由于每一个孤立树在形成的过程中，所随机得到数据属性及与数据属性对应的分裂值不同，这就导致了孤立森林中能包括多个孤立树。孤立树中若设置异常点比例得当，即可提升异常点的检测效果。

通过上述方式，在对多个聚类簇分别构建了孤立森林模型后，每一聚类簇根据其对应的孤立森林模型进行数据分类。

S120、将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心。

在本实施例中，当选定多个聚类簇其中一个聚类簇作为目标聚类簇为示例进行最优异常点比例获取时，需根据初始设置的当前异常点比例将所选定的聚类簇由所述孤立森林模型进行分类后，可以确定分类结果中正常类别的数据点对应的正常点中心，这一正常点中心在后续过程中是恒定不变的。

在一实施例中，如图 4 所示，步骤 S120 包括：

S121、将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到与所选定的聚类簇对应的分类结果；其中，所述分类结果中包括正常类别的数据点和异常类别的数据点；

S122、获取所述分类结果中正常类别的数据点所对应的平均值，以获取初始正常点中心；

S123、获取所述分类结果中正常类别的数据点中与所述初始正常点中心距离最近的数据点，以作为正常类别的数据点对应的正常点中心。

在本实施例中，先根据所述孤立森林模型及当前异常点比例将所选定的聚

类簇进行分类后，得到了包括正常类别的数据点和异常类别的数据点的分类结果。此时为了确定正常点中心，需先获取正常类别的数据点的平均值，然后将正常类别的数据点中距离该平均值最近的数据点，以作为正常点中心。当固定所述正常点中心后，即可不断调整异常点比例，根据指定参数（如当前异常类别的每一数据点与所述正常点中心的平均欧式距离）的变化趋势，来获取最优异常点比例。

S130、获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离。

在本申请中，为了判断异常类别的每一数据点与正常点的距离关系，需计算异常类别的每一数据点与所述正常点中心的欧式距离后求平均，得到所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离，从该当前状态平均欧式距离可以看出异常类别的每一数据点是否均远离正常点中心。

S140、通过所述当前异常点比例减去预设的步长，以更新当前异常点比例。

在本实施例，将所述当前异常点比例减去预设的步长，是为了不断调整当前异常点比例，以通过试探法得出最优异常点比例。

S150、将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离。

在本实施例中，通过将当前异常点比例减去所述步长以更新当前异常点比例，此时无需再次确定正常点中心，只需得到分类结果中的异常类别的数据点，再计算异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离。

S160、通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度。

在本实施例中，通过例如步骤 S130 中得到的当前状态平均欧式距离视为 d_0 ，则步骤 S150 初次执行得到的下一状态平均欧式距离视为 d_1 ，则步骤 S150 第二次执行得到的下一状态平均欧式距离视为 d_2 （此时对应的当前状态平均欧式距离为 d_1 ），……，步骤 S150 第 N 次执行得到的下一状态平均欧式距离视为 d_N （此时对应的当前状态平均欧式距离为 d_{N-1} ）。若将预设的步长记为 l ，则是通

过 $(d_N - d_{N-1}) / l$ 来计算平均欧式距离变动幅度, 其中 N 为大于 0 的正整数。

S170、判断所述平均欧式距离变动幅度是否超出预设的变动幅度阈值。

在本实施例中, 当平均欧式距离变动幅度陡然变大, 表示此刻最新的当前异常点比例不是最优异常点比例, 可考虑将此刻最新的当前异常点比例之前一个状态的当前异常点比例作为最优异常点比例。

S180、若所述平均欧式距离变动幅度超出所述变动幅度阈值, 将当前异常点比例加上步长作为最优异常点比例。

在本实施例中, 若平均欧式距离变动幅度超出预设的变动幅度阈值, 表示有部分真实的异常点被划分为正常点, 导致异常点到正常中心点的平均欧式距离突增, 此时当前异常点比例的上一状态 (即当前异常点比例加上步长) 即可作为最优异常点比例。

S181、将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类, 得到最优分类结果。

在本实施例中, 当确定了最优异常点比例后, 即可将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类, 得到最优分类结果, 得到分类效果较好的无监督分类模型。

在一实施例中, 如图 5 所示, 步骤 S170 之后还包括:

S190、若所述平均欧式距离变动幅度未超出所述变动幅度阈值, 将当前异常点比例减去步长以更新当前异常点比例, 通过下一状态平均欧式距离以更新当前状态平均欧式距离, 返回执行步骤 S150。

在本实施例中, 当平均欧式距离变动幅度仍保持平稳过渡, 表示所降低的异常点比例不足以明显影响异常类别的每一数据点与所述正常点中心的平均欧式距离, 此时需将当前异常点比例减去步长以更新当前异常点比例, 并通过下一状态平均欧式距离以更新新当前状态平均欧式距离。例如当 $(d_N - d_{N-1}) / l$ 未超出预设的变动幅度阈值, 此时将 d_1 作为当前状态平均欧式距离, 将 $(m_0 - 1)$ 作为当前异常点比例重新返回执行步骤 S150 以得到 d_2 ; 之后再次流向步骤 S170 时即是以 $(d_2 - d_1) / l$ 作为平均欧式距离变动幅度, 以此类推, 直至执行到平均欧式距离变动幅度超出预设的变动幅度阈值即可。

在一实施例中, 步骤 S181 之后还包括:

将所述最优分类结果及所述最优异常点比例发送至所述待分类数据点集合

对应的用户端，并将所述最优分类结果及所述最优异常点比例同步发送至云服务器；

将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除。

在本实施例中，若在服务器中完成了获取了与所述待分类数据点集合对应的最优分类结果及所述最优异常点比例后，可以及时的将该最优分类结果及所述最优异常点比例发送至所述待分类数据点集合对应的用户端，实现对用户端进行分类结果的有效通知。

而且为了降低服务器中的数据存储压力，此时可及时的将所述最优分类结果及所述最优异常点比例同步发送至云服务器，通过云服务器实现对与所述待分类数据点集合对应的最优分类结果及所述最优异常点比例的有效存储。此过程中，还可以将与所述最优分类结果及所述最优异常点比例对应的所述待分类数据点集合同步至云服务器。上述的待分类数据点集合、最优分类结果及最优异常点比例在由服务器同步至云服务器中时，需以用户端的唯一机器识别码（如IMEI 串号）为数据标识位来进行唯一数据标识。

此时将所述最优分类结果及所述最优异常点比例同步发送至云服务器之后，则可对服务器中将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除，从而有效释放出存储空间。

在一实施例中，所述将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除之前，还包括：

根据预设的当前异常点比例与所述最优异常点比例之差除以所述步长，得到迭代次数；

将所述迭代次数发送至所述待分类数据点集合对应的用户端，并将所述迭代次数同步发送至云服务器。

在本实施例中，为了清楚的获知预设的当前异常点比例所述最优异常点比例之间经过了多少次迭代，此时可以根据预设的当前异常点比例与所述最优异常点比例之差除以所述步长，得到迭代次数。当获知了所述迭代次数后，可以将所述迭代次数发送至所述待分类数据点集合对应的用户端，用户端对应则可积累设置最优异常点比例的经验。

该方法实现了对海量数据的谱聚类分类，然后分别对各聚类簇同时进行异

常点检测和最优异常点比例自动获取，确定了最优异常点比例后对各聚类簇的数据进行异常点检测，检测准确率得到了提升。

本申请实施例还提供一种基于谱聚类的异常点比例优化装置，该基于谱聚类的异常点比例优化装置用于执行前述基于谱聚类的异常点比例优化方法的任一实施例。具体地，请参阅图 6，图 6 是本申请实施例提供的基于谱聚类的异常点比例优化装置的示意性框图。该基于谱聚类的异常点比例优化装置 100 可以配置于服务器中。

如图 6 所示，基于谱聚类的异常点比例优化装置 100 包括谱聚类单元 101、孤立森林模型训练单元 110、正常点中心获取单元 120、第一平均距离计算单元 130、第一异常点比例更新单元 140、第二平均距离计算单元 150、平均距离变动幅度获取单元 160、幅度判断单元 170、最优比例获取单元 180、最优分类单元 181。

谱聚类单元 101，用于接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇。

在一实施例中，如图 7 所示，谱聚类单元 101 包括：

初始输入单元 1011，用于获取所输入的相似度矩阵和目标聚类数目；

相似度矩阵获取单元 1012，用于根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似矩阵；

拉普拉斯矩阵获取单元 1013，用于根据所述相似矩阵构建邻接矩阵和对角矩阵，由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵；

目标特征向量集获取单元 1014，用于获取所述拉普拉斯矩阵中的多个特征值的排名，若判断特征值的排名位于预设排名阈值之前，获取对应的特征向量以组成目标特征向量集；

目标向量矩阵获取单元 1015，用于将目标特征向量集合中每一特征向量转置为列向量并依次组合，以得到目标向量矩阵；

子图获取单元 1016，用于通过 k-means 算法将目标向量矩阵中各行向量进行聚类，得到与所述目标聚类数目相同的子图，并得到与各子图对应的聚类簇。

孤立森林模型训练单元 110，用于获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型。

在一实施例中，如图 8 所示，孤立森林模型训练单元 110 包括：

分类参数获取单元 111，用于从各聚类簇中均随机获取一个数据属性，及由各聚类簇中所选定的数据属性和当前异常点比例所确定的分裂值；

模型获取单元 112，用于根据所述数据属性及所述分裂值将各聚类簇分别进行划分，得到各聚类簇分别对应的多个孤立树，以组合得到与各聚类簇对应的用于异常点检测的孤立森林模型。

正常点中心获取单元 120，用于将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心。

在一实施例中，如图 9 所示，正常点中心获取单元 120 包括：

初始分类单元 121，用于将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到与所选定的聚类簇对应的分类结果；其中，所述分类结果中包括正常类别的数据点和异常类别的数据点；

距离均值计算单元 122，用于获取所述分类结果中正常类别的数据点所对应的平均值，以获取初始正常点中心；

正常点中心选定单元 123，用于获取所述分类结果中正常类别的数据点中与所述初始正常点中心距离最近的数据点，以作为正常类别的数据点对应的正常点中心。

第一平均距离计算单元 130，用于获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离。

第一异常点比例更新单元 140，用于通过所述当前异常点比例减去预设的步长，以更新当前异常点比例。

第二平均距离计算单元 150，用于将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离。

平均距离变动幅度获取单元 160，用于通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度。

幅度判断单元 170，用于判断所述平均欧式距离变动幅度是否超出预设的变动幅度阈值。

最优比例获取单元 180，用于若所述平均欧式距离变动幅度超出所述变动幅度阈值，将当前异常点比例加上步长作为最优异常点比例。

最优分类单元 181, 用于将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类, 得到最优分类结果。

在一实施例中, 如图 10 所示, 基于谱聚类的异常点比例优化装置 100 还包括:

第二异常点比例更新单元 190, 用于若所述平均欧式距离变动幅度未超出所述变动幅度阈值, 将当前异常点比例减去步长以更新当前异常点比例, 通过下一状态平均欧式距离以更新当前状态平均欧式距离, 返回执行将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类, 得到当前异常类别的数据点, 获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离的步骤。

该装置实现了对海量数据的谱聚类分类, 然后分别对各聚类簇同时进行异常点检测和最优异常点比例自动获取, 确定了最优异常点比例后对各聚类簇的数据进行异常点检测, 检测准确率得到了提升。

上述基于谱聚类的异常点比例优化装置可以实现为计算机程序的形式, 该计算机程序可以在如图 11 所示的计算机设备上运行。

请参阅图 11, 图 11 是本申请实施例提供的计算机设备的示意性框图。该计算机设备 500 是服务器, 服务器可以是独立的服务器, 也可以是多个服务器组成的服务器集群。

参阅图 11, 该计算机设备 500 包括通过系统总线 501 连接的处理器 502、存储器和网络接口 505, 其中, 存储器可以包括非易失性存储介质 503 和内存存储器 504。

该非易失性存储介质 503 可存储操作系统 5031 和计算机程序 5032。该计算机程序 5032 被执行时, 可使得处理器 502 执行基于谱聚类的异常点比例优化方法。

该处理器 502 用于提供计算和控制能力, 支撑整个计算机设备 500 的运行。

该内存存储器 504 为非易失性存储介质 503 中的计算机程序 5032 的运行提供环境, 该计算机程序 5032 被处理器 502 执行时, 可使得处理器 502 执行基于谱聚类的异常点比例优化方法。

该网络接口 505 用于进行网络通信, 如提供数据信息的传输等。本领域技术人员可以理解, 图 11 中示出的结构, 仅仅是与本申请方案相关的部分结构的

框图，并不构成对本申请方案所应用于其上的计算机设备 500 的限定，具体的计算机设备 500 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

其中，所述处理器 502 用于运行存储在存储器中的计算机程序 5032，以实现本申请实施例公开的基于谱聚类的异常点比例优化方法。

本领域技术人员可以理解，图 11 中示出的计算机设备的实施例并不构成对计算机设备具体构成的限定，在其他实施例中，计算机设备可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。例如，在一些实施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 11 所示实施例一致，在此不再赘述。

应当理解，在本申请实施例中，处理器 502 可以是中央处理单元 (Central Processing Unit, CPU)，该处理器 502 还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

在本申请的另一实施例中提供计算机可读存储介质。该计算机可读存储介质可以为非易失性的计算机可读存储介质。该计算机可读存储介质存储有计算机程序，其中计算机程序被处理器执行时实现本申请实施例公开的基于谱聚类的异常点比例优化方法。

所述存储介质为实体的、非瞬时性的存储介质，例如可以是 U 盘、移动硬盘、只读存储器 (Read-Only Memory, ROM)、磁碟或者光盘等各种可以存储程序代码的实体存储介质。

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，上述描述的设备、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1.一种基于谱聚类的异常点比例优化方法，包括：

接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇；

获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型；

将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心；

获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离；

通过所述当前异常点比例减去预设的步长，以更新当前异常点比例；

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离；

通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度；

判断所述平均欧式距离变动幅度超出预设的变动幅度阈值；

若所述平均欧式距离变动幅度超出所述变动幅度阈值，将当前异常点比例加上步长作为最优异常点比例；以及

将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果。

2.根据权利要求1所述的基于谱聚类的异常点比例优化方法，其中，所述通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇，包括：

获取所输入的相似度矩阵和目标聚类数目；

根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似矩阵；

根据所述相似矩阵构建邻接矩阵和对角矩阵，由所述对角矩阵与所述邻接

矩阵之差得到拉普拉斯矩阵；

获取所述拉普拉斯矩阵中的多个特征值的排名，若判断特征值的排名位于预设排名阈值之前，获取对应的特征向量以组成目标特征向量集；

将目标特征向量集合中每一特征向量转置为列向量并依次组合，以得到目标向量矩阵；

通过 k-means 算法将目标向量矩阵中各行向量进行聚类，得到与所述目标聚类数目相同的子图，并得到与各子图对应的聚类簇。

3. 根据权利要求 1 所述的基于谱聚类的异常点比例优化方法，其中，所述判断所述平均欧式距离变动幅度超出预设的变动幅度阈值之后，还包括：

若所述平均欧式距离变动幅度未超出所述变动幅度阈值，将当前异常点比例减去步长以更新当前异常点比例，通过下一状态平均欧式距离以更新当前状态平均欧式距离，返回执行将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离的步骤。

4. 根据权利要求 1 所述的基于谱聚类的异常点比例优化方法，其中，所述根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型，包括：

从各聚类簇中均随机获取一个数据属性，及由各聚类簇中所选定的数据属性和当前异常点比例所确定的分裂值；

根据所述数据属性及所述分裂值将各聚类簇分别进行划分，得到各聚类簇分别对应的多个孤立树，以组合得到与各聚类簇对应的用于异常点检测的孤立森林模型。

5. 根据权利要求 1 所述的基于谱聚类的异常点比例优化方法，其中，所述将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心，包括：

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到与所选定的聚类簇对应的分类结果；其中，所述分类结果中包括正常类别的数据点和异常类别的数据点；

获取所述分类结果中正常类别的数据点所对应的平均值，以获取初始正常点中心；

获取所述分类结果中正常类别的数据点中与所述初始正常点中心距离最近的数据点，以作为正常类别的数据点对应的正常点中心。

6. 根据权利要求 1 所述的基于谱聚类的异常点比例优化方法，其中，所述将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果之后，还包括：

将所述最优分类结果及所述最优异常点比例发送至所述待分类数据点集合对应的用户端，并将所述最优分类结果及所述最优异常点比例同步发送至云服务器；

将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除。

7. 根据权利要求 6 所述的基于谱聚类的异常点比例优化方法，其中，所述将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除之前，还包括：

根据预设的当前异常点比例与所述最优异常点比例之差除以所述步长，得到迭代次数；

将所述迭代次数发送至所述待分类数据点集合对应的用户端，并将所述迭代次数同步发送至云服务器。

8. 一种基于谱聚类的异常点比例优化装置，包括：

谱聚类单元，用于接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇；

孤立森林模型训练单元，用于获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型；

正常点中心获取单元，用于将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心；

第一平均距离计算单元，用于获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离；

第一异常点比例更新单元，用于通过所述当前异常点比例减去预设的步长，以更新当前异常点比例；

第二平均距离计算单元，用于将所选定的聚类簇根据所述孤立森林模型及

当前异常点比例进行分类,得到当前异常类别的数据点,获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离;

平均距离变动幅度获取单元,用于通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长,得到平均欧式距离变动幅度;

幅度判断单元,用于判断所述平均欧式距离变动幅度超出预设的变动幅度阈值;

最优比例获取单元,用于若所述平均欧式距离变动幅度超出所述变动幅度阈值,将当前异常点比例加上步长作为最优异常点比例;以及

最优分类单元,用于将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类,得到最优分类结果。

9.根据权利要求8所述的基于谱聚类的异常点比例优化装置,其中,所述谱聚类单元,包括:

初始输入单元,用于获取所输入的相似度矩阵和目标聚类数目;

相似度矩阵获取单元,用于根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似矩阵;

拉普拉斯矩阵获取单元,用于根据所述相似矩阵构建邻接矩阵和对角矩阵,由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵;

目标特征向量集获取单元,用于获取所述拉普拉斯矩阵中的多个特征值的排名,若判断特征值的排名位于预设排名阈值之前,获取对应的特征向量以组成目标特征向量集;

目标向量矩阵获取单元,用于将目标特征向量集合中每一特征向量转置为列向量并依次组合,以得到目标向量矩阵;

子团获取单元,用于通过k-means算法将目标向量矩阵中各行向量进行聚类,得到与所述目标聚类数目相同的子图,并得到与各子图对应的聚类簇。

10.根据权利要求8所述的基于谱聚类的异常点比例优化装置,其中,还包括:

第二异常点比例更新单元,用于若所述平均欧式距离变动幅度未超出所述变动幅度阈值,将当前异常点比例减去步长以更新当前异常点比例,通过下一状态平均欧式距离以更新当前状态平均欧式距离,返回执行将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类,得到当前异常类别的数据

点, 获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离的步骤。

11. 一种计算机设备, 包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序, 所述处理器执行所述计算机程序时实现以下步骤:

接收待分类数据点集合, 通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图, 并得到与各子图对应的聚类簇;

获取多个聚类簇中所包括每一聚类簇对应的数据点, 根据预设的当前异常点比例及每一聚类簇, 构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型;

将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类, 得到分类结果中正常类别的正常点中心;

获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离, 以作为当前状态平均欧式距离;

通过所述当前异常点比例减去预设的步长, 以更新当前异常点比例;

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类, 得到当前异常类别的数据点, 获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离;

通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长, 得到平均欧式距离变动幅度;

判断所述平均欧式距离变动幅度超出预设的变动幅度阈值;

若所述平均欧式距离变动幅度超出所述变动幅度阈值, 将当前异常点比例加上步长作为最优异常点比例; 以及

将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类, 得到最优分类结果。

12. 根据权利要求 11 所述的计算机设备, 其中, 所述通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图, 并得到与各子图对应的聚类簇, 包括:

获取所输入的相似度矩阵和目标聚类数目;

根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似

矩阵；

根据所述相似矩阵构建邻接矩阵和对角矩阵，由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵；

获取所述拉普拉斯矩阵中的多个特征值的排名，若判断特征值的排名位于预设排名阈值之前，获取对应的特征向量以组成目标特征向量集；

将目标特征向量集合中每一特征向量转置为列向量并依次组合，以得到目标向量矩阵；

通过 k-means 算法将目标向量矩阵中各行向量进行聚类，得到与所述目标聚类数目相同的子图，并得到与各子图对应的聚类簇。

13. 根据权利要求 11 所述的计算机设备，其中，所述判断所述平均欧式距离变动幅度超出预设的变动幅度阈值之后，还包括：

若所述平均欧式距离变动幅度未超出所述变动幅度阈值，将当前异常点比例减去步长以更新当前异常点比例，通过下一状态平均欧式距离以更新当前状态平均欧式距离，返回执行将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离的步骤。

14. 根据权利要求 11 所述的计算机设备，其中，所述根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型，包括：

从各聚类簇中均随机获取一个数据属性，及由各聚类簇中所选定的数据属性和当前异常点比例所确定的分裂值；

根据所述数据属性及所述分裂值将各聚类簇分别进行划分，得到各聚类簇分别对应的多个孤立树，以组合得到与各聚类簇对应的用于异常点检测的孤立森林模型。

15. 根据权利要求 11 所述的计算机设备，其中，所述将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心，包括：

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到与所选定的聚类簇对应的分类结果；其中，所述分类结果中包括正常类别的数据点和异常类别的数据点；

获取所述分类结果中正常类别的数据点所对应的平均值，以获取初始正常点中心；

获取所述分类结果中正常类别的数据点中与所述初始正常点中心距离最近的数据点，以作为正常类别的数据点对应的正常点中心。

16. 根据权利要求 11 所述的计算机设备，其中，所述将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果之后，还包括：

将所述最优分类结果及所述最优异常点比例发送至所述待分类数据点集合对应的用户端，并将所述最优分类结果及所述最优异常点比例同步发送至云服务器；

将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除。

17. 根据权利要求 16 所述的计算机设备，其中，所述将所述最优分类结果及所述最优异常点比例对应的存储区域进行格式化删除之前，还包括：

根据预设的当前异常点比例与所述最优异常点比例之差除以所述步长，得到迭代次数；

将所述迭代次数发送至所述待分类数据点集合对应的用户端，并将所述迭代次数同步发送至云服务器。

18. 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序当被处理器执行时使所述处理器执行以下操作：

接收待分类数据点集合，通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇；

获取多个聚类簇中所包括每一聚类簇对应的数据点，根据预设的当前异常点比例及每一聚类簇，构建与每一聚类簇一一对应的用于异常点检测的孤立森林模型；

将所选定的聚类簇根据所述孤立森林模型及所述当前异常点比例进行分类，得到分类结果中正常类别的正常点中心；

获取所述分类结果中异常类别的每一数据点与所述正常点中心的平均欧式距离，以作为当前状态平均欧式距离；

通过所述当前异常点比例减去预设的步长，以更新当前异常点比例；

将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离；

通过下一状态平均欧式距离与当前状态平均欧式距离之差除以所述步长，得到平均欧式距离变动幅度；

判断所述平均欧式距离变动幅度超出预设的变动幅度阈值；

若所述平均欧式距离变动幅度超出所述变动幅度阈值，将当前异常点比例加上步长作为最优异常点比例；以及

将所选定的聚类簇根据所述孤立森林模型及最优异常点比例进行分类，得到最优分类结果。

19. 根据权利要求 18 所述的计算机可读存储介质，其中，所述通过谱聚类将所述待分类数据点集合对应的节点划分为多个子图，并得到与各子图对应的聚类簇，包括：

获取所输入的相似度矩阵和目标聚类数目；

根据所述相似度矩阵构建与所述待分类数据点集合对应的节点相应的相似矩阵；

根据所述相似矩阵构建邻接矩阵和对角矩阵，由所述对角矩阵与所述邻接矩阵之差得到拉普拉斯矩阵；

获取所述拉普拉斯矩阵中的多个特征值的排名，若判断特征值的排名位于预设排名阈值之前，获取对应的特征向量以组成目标特征向量集；

将目标特征向量集合中每一特征向量转置为列向量并依次组合，以得到目标向量矩阵；

通过 k-means 算法将目标向量矩阵中各行向量进行聚类，得到与所述目标聚类数目相同的子图，并得到与各子图对应的聚类簇。

20. 根据权利要求 18 所述的计算机可读存储介质，其中，所述判断所述平均欧式距离变动幅度超出预设的变动幅度阈值之后，还包括：

若所述平均欧式距离变动幅度未超出所述变动幅度阈值，将当前异常点比例减去步长以更新当前异常点比例，通过下一状态平均欧式距离以更新当前状态平均欧式距离，返回执行将所选定的聚类簇根据所述孤立森林模型及当前异常点比例进行分类，得到当前异常类别的数据点，获取当前异常类别的每一数

据点与所述正常点中心的平均欧式距离以作为下一状态平均欧式距离的步骤。

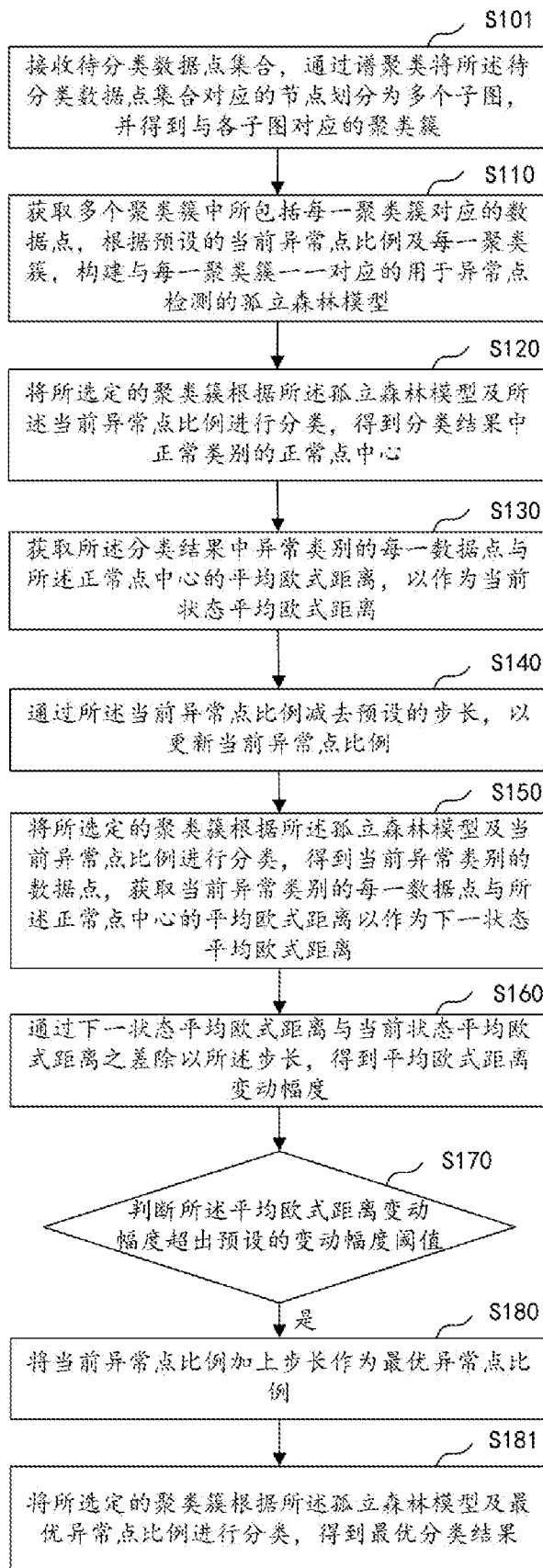


图 1

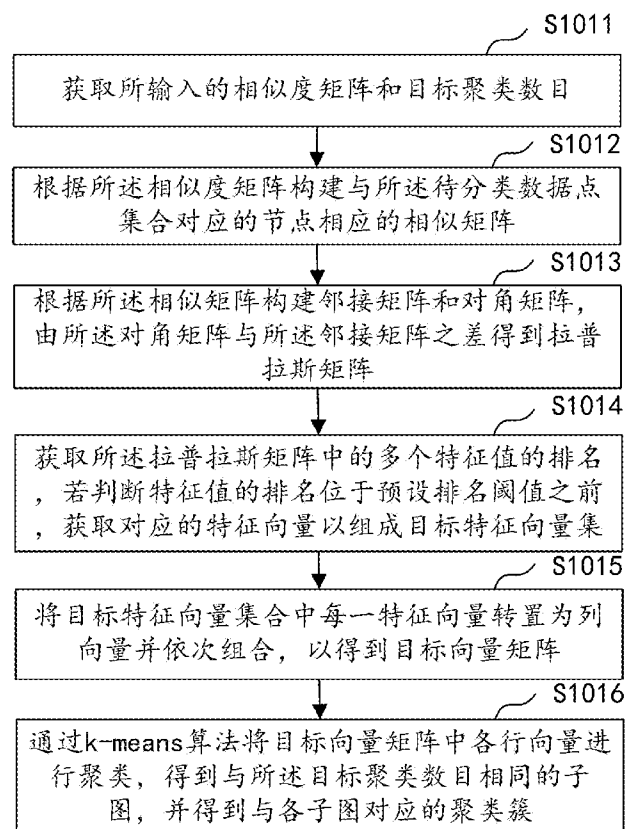


图 2

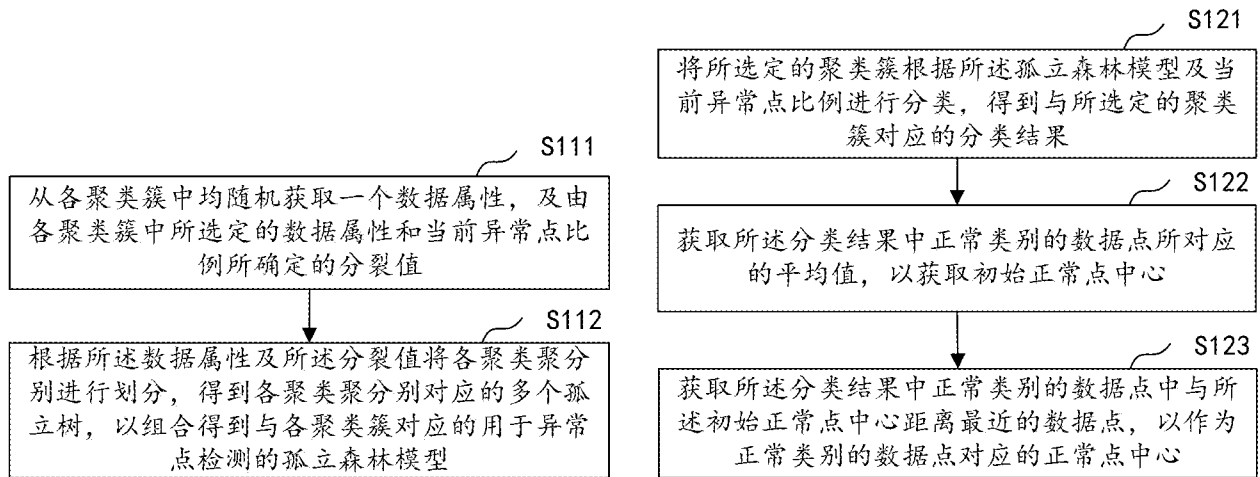


图 3

图 4

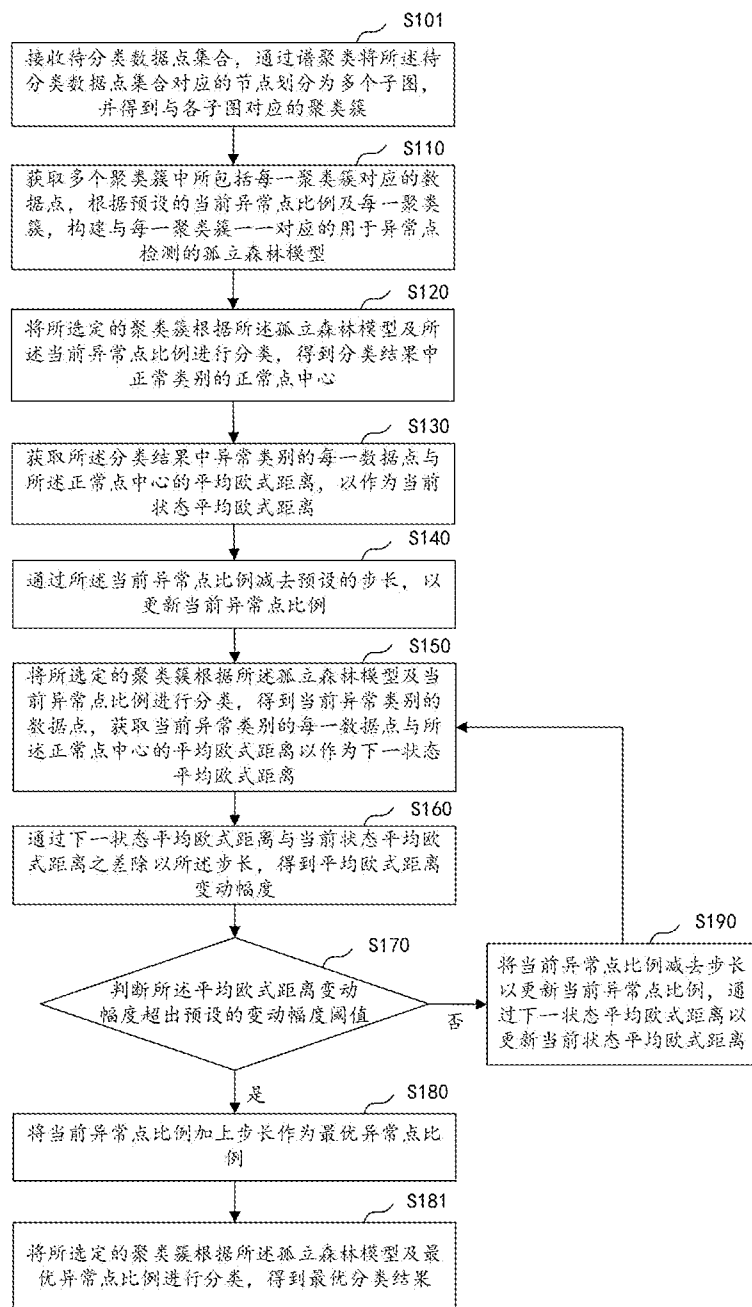


图 5

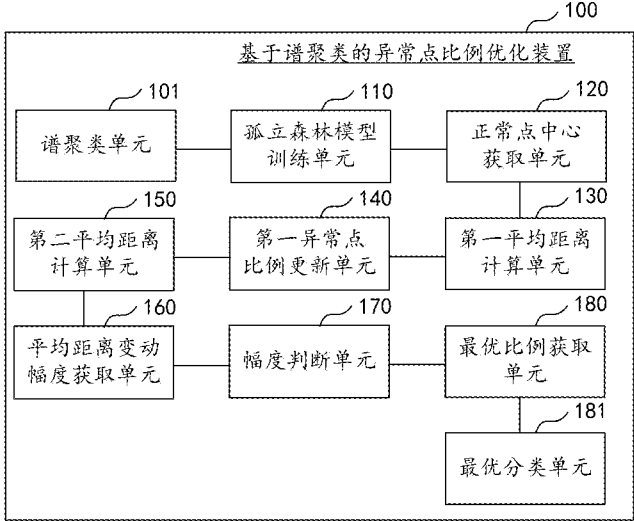


图 6

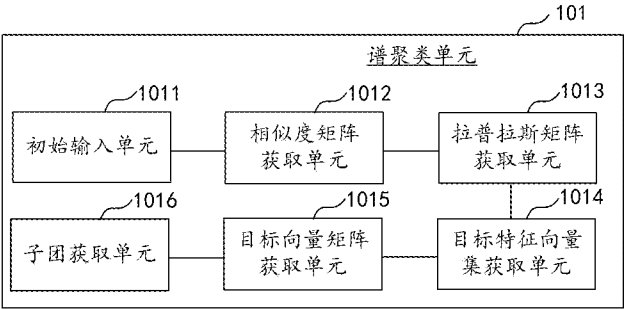


图 7

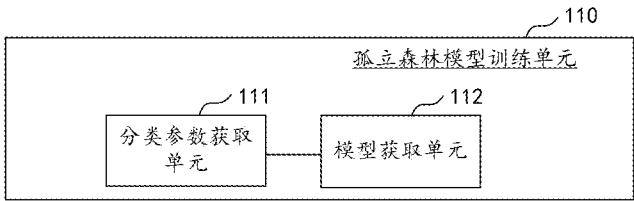


图 8

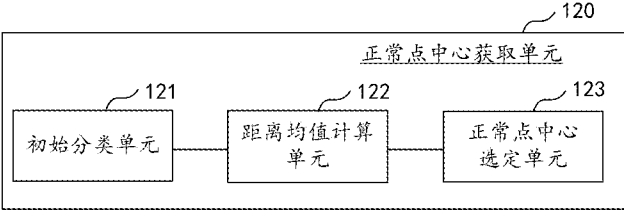


图 9

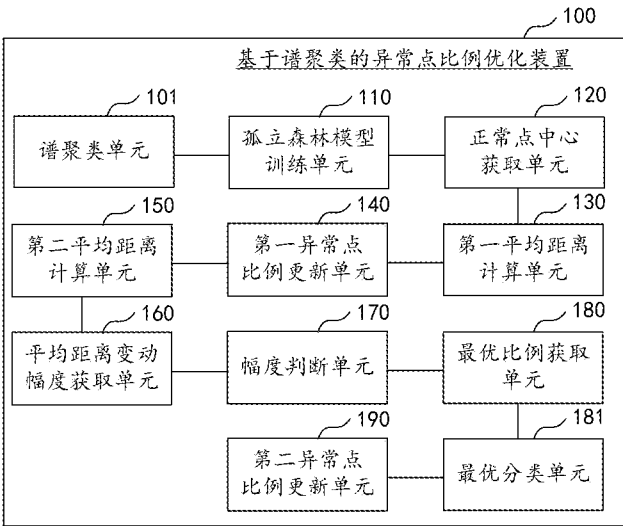


图 10

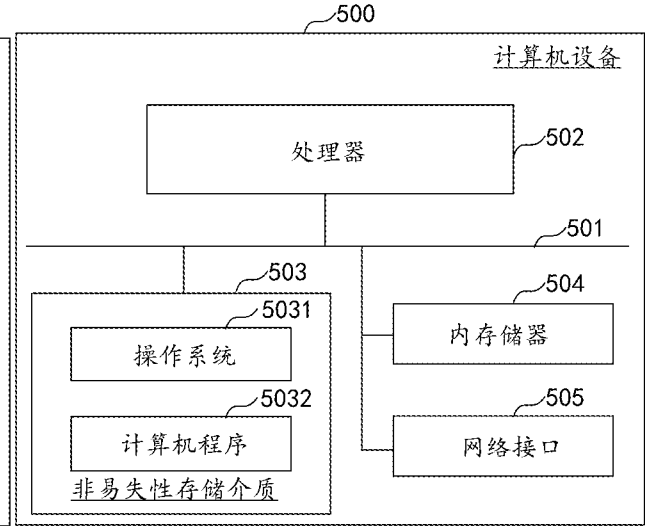


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117355

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K;G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI; SIPOABS; CNABS; CNKI: 异常, 孤立森林, 聚类, 谱, 数据, 比例, 分类, abnormal, cluster, data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109871886 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 11 June 2019 (2019-06-11) entire document	1-20
A	CN 108322363 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 24 July 2018 (2018-07-24) entire document	1-20
A	CN 109145934 A (BEIJING SHU'AN XINYUN INFORMATION TECHNOLOGY CO., LTD.) 04 January 2019 (2019-01-04) entire document	1-20
A	CN 108475250 A (HUAWEI TECHNOLOGIES CO., LTD.) 31 August 2018 (2018-08-31) entire document	1-20
A	WO 2019003703 A1 (HITACHI, LTD.) 03 January 2019 (2019-01-03) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

05 February 2020

Date of mailing of the international search report

18 February 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/117355

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109871886	A	11 June 2019	None			
CN	108322363	A	24 July 2018	None			
CN	109145934	A	04 January 2019	CN	109145934	B	21 May 2019
				WO	2019120241	A1	27 June 2019
CN	108475250	A	31 August 2018	US	10193780	B2	29 January 2019
				WO	2017059807	A1	13 April 2017
				US	2017104657	A1	13 April 2017
WO	2019003703	A1	03 January 2019	CN	110506245	A	26 November 2019
				JP	2019008681	A	17 January 2019

国际检索报告

国际申请号

PCT/CN2019/117355

A. 主题的分类

G06K 9/62 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

DWPI; SIPOABS; CNABS; CNKI:异常, 孤立森林, 聚类, 谱, 数据, 比例, 分类, abnormal, cluster, data,

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 109871886 A (平安科技深圳有限公司) 2019年 6月 11日 (2019 - 06 - 11) 全文	1-20
A	CN 108322363 A (腾讯科技深圳有限公司) 2018年 7月 24日 (2018 - 07 - 24) 全文	1-20
A	CN 109145934 A (北京数安鑫云信息技术有限公司) 2019年 1月 4日 (2019 - 01 - 04) 全文	1-20
A	CN 108475250 A (华为技术有限公司) 2018年 8月 31日 (2018 - 08 - 31) 全文	1-20
A	W0 2019003703 A1 (HITACHI LTD) 2019年 1月 3日 (2019 - 01 - 03) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 2月 5日

国际检索报告邮寄日期

2020年 2月 18日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

明媚

电话号码 62411851

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117355

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	109871886	A	2019年 6月 11日	无	
CN	108322363	A	2018年 7月 24日	无	
CN	109145934	A	2019年 1月 4日	CN	109145934 B 2019年 5月 21日
				WO	2019120241 A1 2019年 6月 27日
CN	108475250	A	2018年 8月 31日	US	10193780 B2 2019年 1月 29日
				WO	2017059807 A1 2017年 4月 13日
				US	2017104657 A1 2017年 4月 13日
WO	2019003703	A1	2019年 1月 3日	CN	110506245 A 2019年 11月 26日
				JP	2019008681 A 2019年 1月 17日