

疾病异常数据检测方法、计算机装置及存储介质

本申请要求于2018年04月11日提交中国专利局,申请号为201810321852.9发明名称为“疾病异常数据检测方法、计算机装置及存储介质”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

本申请涉及数据处理技术领域,具体涉及一种疾病异常数据检测方法、计算机装置和非易失性可读存储介质。

背景技术

随着全球经济一体化进程的加快,经济与交流活动增加,人群流动日益频繁,为疾病的传播与爆发提供了有利环境,公共卫生健康问题越来越严峻。同时,社会与自然环境也发生着变化,环境污染、自然灾害等影响公众健康事件的增多也增加了突发公共卫生事件爆发的可能性。

如何检测出疾病异常数据,以便能够早期识别到疾病流行或爆发的突发公共卫生事件,尽早采取相应的控制措施,将突发公共卫生事件造成的损失降到最低,成为亟待解决的问题。

现有的异常检测方法,例如 zscore 异常检测方法、格拉布斯(Grubbs)异常检测方法都要求数据满足正态分布,而实际上很多时候无法满足这个要求。对于传统的四分位法,会用到所有的数据,而过去久远的数据对当前数据的参考价值低,结果更容易出现偏差。

发明内容

鉴于以上内容,有必要提出一种疾病异常数据检测方法、计算机装置和非易失性可读存储介质,其可以实现高效准确的疾病监测数据异常检测。

本申请的第一方面提供一种疾病异常数据检测方法,所述方法包括:

获取时间点 0 到时间点 t 的疾病监测数据,构成时间序列数据 X, $X = [x_0, x_1, x_2, \dots, x_t]$;

选取时间窗大小 w, 针对时间点 w 至时间点 t 中的每个时间点 i, 计算所述时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 所述时间点 i 对应的时间窗的大小为 w, $i = w, w + 1, \dots, t$;

根据每个时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 计算

所述时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i , 得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$, 其中

$$mz_i = \frac{|x_i - \mu_i|}{\sigma_i};$$

根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值;

判断时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值, 若时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值, 则所述时间点*i*对应的疾病监测数据 x_i 为异常值。

本申请的第二方面提供一种疾病异常数据检测装置, 所述装置包括:

获取单元, 用于获取时间点 0 到时间点 t 的疾病监测数据, 构成时间序列数据 X, $X = [x_0, x_1, x_2, \dots, x_t]$;

计算单元, 用于选取时间窗大小 w, 针对时间点 w 至时间点 t 中的每个时间点*i*, 计算所述时间点*i*对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 所述时间点*i*对应的的时间窗的大小为 w, $i = w, w + 1, \dots, t$;

所述计算单元, 还用于根据每个时间点*i*对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 计算所述时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i , 得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$, 其中

$$mz_i = \frac{|x_i - \mu_i|}{\sigma_i};$$

确定单元, 用于根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值;

判断单元, 用于判断时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值, 若时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值, 则所述时间点*i*对应的疾病监测数据 x_i 为异常值。

本申请的第三方面提供一种计算机装置, 所述计算机装置包括存储器和处理器, 所述存储器存储至少一条计算机可读指令, 所述处理器执行所述至少一条计算机可读指令以实现所述疾病异常数据检测方法。

本申请的第四方面提供一种非易失性可读存储介质, 所述非易失性可读存储介质上存储有至少一条计算机可读指令, 其特征在于, 所述至少一条计算机可读指令被处理器执行时实现所述疾病异常数据检测方法。

本申请获取时间点 0 到时间点 t 的疾病监测数据, 构成时间序列数据 X, $X = [x_0, x_1, x_2, \dots, x_t]$; 选取时间窗大小 w, 针对时间点 w 至时间点 t 中的每个时间点*i*, 计算所述时间点*i*对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 所述时间点*i*对应的的时间窗的大小为 w, $i = w, w + 1, \dots, t$; 根据每个时间点*i*对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 计算所述时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i , 得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$; 根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值; 判断时间点*i*对应的疾病监测数据 x_i 的 moving-zscore 值是否大

于所述异常值阈值，若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值，则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

本申请引入了时间窗，将最具有参考价值的历史邻近数据分布来作为参考，检测结果更加合理。由于时间窗内数据量较少，导致时间窗内的数据很难满足正态分布，因而不能采用传统的 zscore 方法检测疾病监测数据中的异常值。本申请计算多个疾病监测数据的 moving-zscore 值，保证了数量足够大，能够满足正态分布的条件，然后再根据所有 moving-zscore 值确定异常值阈值，从而检测出疾病监测数据中的异常值。因此，本申请对于疾病监测数据的异常检测具有更高的可适用性，实现了高效准确的疾病监测数据异常检测。

附图说明

图 1 是本申请实施例一提供的疾病异常数据检测方法的流程图。

图 2 是本申请实施例二提供的疾病异常数据检测装置的结构图。

图 3 是本申请实施例四提供的计算机装置的示意图。

具体实施方式

为了能够更清楚地理解本申请的上述目的、特征和优点，下面结合附图和具体实施例对本申请进行详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

在下面的描述中阐述了很多具体细节以便于充分理解本申请，所描述的实施例仅仅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。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在限制本申请。

优选地，本申请的疾病异常数据检测方法应用在一个或者多个计算机装置中。所述计算机装置是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit, ASIC)、可编程门阵列(Field Programmable Gate Array, FPGA)、数字处理器(Digital Signal Processor, DSP)、嵌入式设备等。

所述计算机装置可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机装置可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

实施例一

图 1 是本申请实施例一提供的疾病异常数据检测方法的流程图。所述疾病异常数据检测方法应用于计算机装置。所述疾病异常数据检测方法检测疾病监测数据中的异常值，以便能够早期识别到疾病流行或爆发的突发公共卫生事件，尽早采取相应的控制措施，将突发公共卫生事件造成的损失降到最低。

如图 1 所示，所述疾病异常数据检测方法具体包括以下步骤：

步骤 101，获取时间点 0 到时间点 t 的疾病监测数据，构成时间序列数据 X ， $X = [x_0, x_1, x_2, \dots, x_t]$ 。

所述疾病监测数据可以包括流感、手足口病、麻疹、流行性腮腺炎等疾病的监测数据。

可以在预设区域（例如省市、地区）建立由多个监测点组成的疾病监测网络，从所述监测点获取所述疾病监测数据。可以选择医疗机构、学校和幼托机构、药店等作为监测点，分别对相应的目标人群进行疾病监测及数据采集。可以选择满足预设条件的场所作为监测点。所述预设条件可以包括人数、规模等。例如，选择学生人数达到预设数量的学校和幼托机构作为监控点。又如，选择规模（例如以日营业额统计）达到预设规模的药店作为监控点。再如，选择规模（例如以日就医人数统计）达到预设规模的医院作为监控点。

不同时间的疾病监测数据构成疾病监测的时间序列数据。例如，可以将以日为单位采集到的疾病监测数据构成疾病监测的时间序列数据。或者，可以将以周为单位采集到的疾病监测数据构成疾病监测的时间序列数据。

医疗机构（主要包括医院）是最能捕捉疾病早期暴发预兆的场所，是开展疾病监测的首选。可以根据病人就诊情况，获取疾病监测数据。

一部分病人会自行去药店购药来缓解早期症状，因此，可以根据药店的药品销售情况，获取疾病监测数据。

儿童和青少年是疾病的高危人群以及疾病传播过程中的重要环节，也应该加强对该人群的监测。学校和幼托机构是监测儿童和青少年疾病发病情况的较佳场所。可以根据学校和幼托机构的儿童和青少年的请假情况，获得疾病监测数据。

因此，本申请中主要选择医疗机构、学校和幼托机构、药店这三类场所进行疾病监测数据的采集。当然，上述对数据源的选择，并不能限制在另外的实施方案中增加或替换其他重点关注人群或场所作为监测的数据源。例如，可以将宾馆纳入疾病监测范围，获取宾馆入住人员的疾病监测数据。

根据需要，可以取任意一类监控点（例如医疗机构）采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以取医院采集的疾病监测数据构成疾病监测的时间序列数据。或者，可以结合多类监控点采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以以医院采集的疾病监测数据为

主，以药店参加的疾病监测数据作为补充，构成疾病监测的时间序列数据。

疾病监测数据可以包括疾病的就诊数、就诊率、发病数、发病率等患病数据。例如，可以从医疗机构（例如医院）获取疾病（例如流感）的每日就诊数，将疾病（例如流感）的每日就诊数作为疾病监测数据。又如，可以从学校获取学生的疾病（例如流感）的每日发病数，将疾病（例如流感）的每日发病数作为疾病监测数据。

需要说明的是，时间点 0 和时间点 t 表示时间序列数据的起始时间和终止时间，并非对时间序列数据时间上的限定。可以将任意时间点作为时间点 0。

步骤 102，选取时间窗大小 w ，针对时间点 w 至时间点 t 中的每个时间点 i ，计算所述时间点 i 对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，所述时间点 i 对应的的时间窗的大小为 w ， $i = w, w + 1, \dots, t$ 。

时间窗为给定时间点之前邻近的一段时间。例如，设时间窗大小 w 为 4，给定时间点为 $t=10$ ，时间窗就是 $t=6$ 到 $t=9$ 这段时间；给定时间点为 $t=11$ ，时间窗就是 $t=7$ 到 $t=11$ 这段时间；依此类推。

时间窗大小 w 是可调参数，可以根据实际情况进行调节。

在一实施例中，所述疾病监测数据为每日的疾病监测数据（例如疾病的每日就诊数），所述时间窗大小 w 可以取为 7（即一周）。

在另一实施例中，所述疾病监测数据为每周的疾病监测数据（例如疾病的每周就诊数），所述时间窗大小 w 可以取为 3（即三周）。

步骤 103，根据每个时间点 i 对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，计算所述时间点 i 对应的的疾病监测数据 x_i 的 moving-zscore 值 mz_i ，得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$ ，其中

$$mz_i = \frac{|x_i - \mu_i|}{\sigma_i}。$$

步骤 102-103 从时间点 w 开始计算各个时间点对应的的时间窗内疾病监测数据的均值和标准差，以及从时间点 w 开始计算各个时间点对应的的疾病监测数据的 moving-zscore 值。这是因为从 $t=0$ 到 $t=w-1$ 这段时间内得到的时间窗的长度小于 w ，没有足够的 w 个疾病监测数据来计算疾病监测数据的均值和标准差。

步骤 104，根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值。

可以计算所述 mz 列表的预设分位数或预设百分位数，将所述预设分位数或预设百分位数作为所述异常值阈值。例如，计算 mz 列表的第 0.997 分位数，将 mz 列表的第 0.997 分位数作为异常值阈值。又如，计算 mz 列表的第 0.5 分位数（即中位数），将 mz 列表的第 0.5 分位数（即中位数）作为异常值阈值。

或者，可以计算 mz 列表的平均数，将所述平均数作为异常值阈值。

步骤 105，判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值，若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值，则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

例如，异常值阈值为 12，时间点 9 对应的疾病监测数据 x_9 的 moving-zscore 值为 13，大于所述异常值阈值，则疾病监测数据 x_9 为异常值。

举例说明：

29 个连续时间点的疾病监测数据为：

[3,3,4,5,4,6,7,6,7,8,9,8,17,8,7,9,10,11,12,12,13,12,14,14,15,16,16,17,16]。

取时间窗大小 $w=4$ ，也就是从第 5 个数 4 开始计算 moving-zscore 值，对于第 5 个数，其时间窗内的数为 [3,3,4,5]，均值为 3.75，标准差为 0.829，则 moving-zscore 值为 $(4-3.75)/0.829=0.302$ （四舍五入保留小数点后三位）。

按照同样的方法，依次计算出第 5 个数后面所有的 moving-zscore 值，得到 mz 列表：

[0.302,2.828,2.713,0.447,1.147,3.0,2.828,0.447,12.728,0.662,0.927,0.246,0.063,2.236,1.859,1.342,2.110,0.0,4.041,1.508,2.110,2.065,1.508,2.111,0.0]。

再取 mz 列表的 0.995 分位数作为异常值阈值，得到 11.685。

列表中只有 12.728 是大于这个异常值阈值的，12.728 对应的疾病监测数据 17，所以 17 就是异常值。

实施例一的疾病异常数据检测方法获取时间点 0 到时间点 t 的疾病监测数据，构成时间序列数据 X ， $X = [x_0, x_1, x_2, \dots, x_t]$ ；选取时间窗大小 w ，针对时间点 w 至时间点 t 中的每个时间点 i ，计算所述时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，所述时间点 i 对应的时间窗的大小为 w ， $i = w, w+1, \dots, t$ ；根据每个时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，计算所述时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i ，得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$ ；根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值；判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值，若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值，则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

实施例一引入了时间窗，将最具有参考价值的历史邻近数据分布来作为参考，检测结果更加合理。由于时间窗内数据量较少，导致时间窗内的数据很难满足正态分布，因而不能采用传统的 zscore 方法检测疾病监测数据中的异常值。实施例一计算多个疾病监测数据的 moving-zscore 值，保证了数量足够大，能够满足正态分布的条件，然后再根据所有 moving-zscore 值确定异常值阈值，从而检测出疾病监测数据中的异常值。因此，实施例一对于疾病监

测数据的异常检测具有更高的可适用性，实现了高效准确的疾病监测数据异常检测。

在另一实施例中，可以结合其他的疾病异常数据检测方法（不同于实施例一的疾病异常数据检测方法）进行疾病异常数据检测。具体地，还可以包括以下步骤：获取其他的疾病异常数据检测方法对所述时间序列数据进行检测得到的异常值；将实施例一的疾病异常数据检测方法得到的异常值与其他的疾病异常数据检测方法得到的异常值进行比较；根据实施例一的疾病异常数据检测方法得到的异常值与其他的疾病异常数据检测方法得到的异常值的比较结果得到最终的异常值。其他的疾病异常数据检测方法可以包括一种，也可以包括多种。每种其他的疾病异常数据检测方法可以得到一组对应的异常值。

例如，采用实施例一的疾病异常数据检测方法对疾病监测的时间序列数据进行疾病异常数据检测得到第一异常值，采用与实施例一的疾病异常数据检测方法不同的第二疾病异常数据检测方法（例如 Grubbs 检测方法）对所述疾病监测的时间序列数据进行疾病异常数据检测得到第二异常值，比较所述第一异常值和所述第二异常值得到最终的异常值。若所述第一异常值和所述第二异常值一致，则以所述第一异常值/到第二异常值作为最终的异常值；否则，若所述第一异常值和所述第二异常值不一致，则舍弃所述第一异常值与所述第二异常值。

又如，采用实施例一的疾病异常数据检测方法对疾病监测的时间序列数据进行疾病异常数据检测得到第一异常值，采用与实施例一的疾病异常数据检测方法不同的第二疾病异常数据检测方法（例如 Grubbs 检测方法）对所述疾病监测的时间序列数据进行疾病异常数据检测得到第二异常值，采用与实施例一的疾病异常数据检测方法不同的第三疾病异常数据检测方法（例如贝叶斯检测方法）对所述疾病监测的时间序列数据进行疾病异常数据检测得到第三异常值，比较所述第一异常值、所述第二异常值和所述第三异常值得到最终的异常值。若所述第一异常值、所述第二异常值和所述第三异常值中至少两个异常值一致，则以一致的异常值作为最终的异常值。

在实施例一的疾病异常数据检测方法的基础上，结合其他的疾病异常数据检测方法进行疾病异常数据检测，可以获得更加准确的疾病异常数据检测结果。

在另一实施例中，可以采用实施例一的疾病异常数据检测方法分别对地区和地区下属医院进行疾病异常数据检测，得到针对地区的检测结果（即地区异常值）和针对地区下属医院的检测结果（即地区下属医院异常值），比较所述地区异常值和所述地区下属医院异常值，根据所述地区异常值和所述地区下属医院异常值的比较结果得到最终的异常值。若针对地区的检测结果

和针对地区下属医院的检测结果一致，则以针对地区的检测结果/针对地区下属医院的检测结果作为最终的检测结果；否则，若针对地区的检测结果和针对地区下属医院的检测结果不一致，则舍弃所述针对地区的检测结果和针对地区下属医院的检测结果。

其中，对地区进行疾病异常数据检测是根据所述地区的疾病监测的时间序列数据（例如地区的疾控部门从各个地区下属医院收集到的疾病监测数据）得到地区异常值，对地区下属医院进行疾病异常数据检测是根据所述地区下属医院的疾病监测的时间序列数据得到所述地区下属医院的异常值。

具体来说，按照实施例一的疾病异常数据检测方法，从地区某时间段内的疾病监测数据（例如病例数）中找出异常值，从该地区下属医院同样时间段内的疾病监测数据中找出异常值，取两个维度的交集来作为最终的检测结果。

例如，地区在 2014-3-3，2014-3-4，2014-3-5，2014-3-6 病例数分别为 160,250，170,180，该地区下属医院（例如该地区下该疾病累计就诊次数最多的医院）在这段时间的病例数分别为 130,180,125，140。从地区病例数分布来看，在 2014-3-4 这天出现了异常病例数，同时下属医院在这天也出现了异常病例数，所以综合判定在这天出现了疾病就诊异常的现象。

又如，地区在 2014-3-3，2014-3-4，2014-3-5，2014-3-6 病例数分别为 160,210，170,180，该地区下属医院（例如该地区下该疾病累计就诊次数最多的医院）在这段时间病例数分别为 130,140,125，140。从地区病例数分布来看，在 2014-3-4 这天出现了异常病例数，但是下属医院在这天并未出现异常病例数。这是由于地区下有很多家医院，地区在这一天的就诊病例数是由下属所有的医院的病例数的总和，部分医院在这一天有一个小幅度的增长，在医院维度上并不是异常，但是所有的小幅度增长后的总和在地区维度上就表现为异常了。

采用实施例一的疾病异常数据检测方法分别对地区和地区下属医院进行疾病异常数据检测，根据针对地区的检测结果和针对地区下属医院的检测结果得到最终的检测结果，可以获得更加准确的疾病异常数据检测结果。

实施例二

图 2 为本申请实施例二提供的疾病异常数据检测装置的结构图。如图 2 所示，所述疾病异常数据检测装置 10 可以包括：获取单元 201、计算单元 202、确定单元 203、判断单元 204。

获取单元 201，用于获取时间点 0 到时间点 t 的疾病监测数据，构成时间序列数据 X ， $X = [x_0, x_1, x_2, \dots, x_t]$ 。

所述疾病监测数据可以包括流感、手足口病、麻疹、流行性腮腺炎等疾病的监测数据。

可以在预设区域（例如省市、地区）建立由多个监测点组成的疾病监测网络，从所述监测点获取所述疾病监测数据，由所述疾病监测数据构成疾病监测的时间序列数据。可以选择医疗机构、学校和幼托机构、药店等作为监测点，分别对相应的目标人群进行疾病监测及数据采集。可以选择满足预设条件的场所作为监测点。所述预设条件可以包括人数、规模等。例如，选择学生人数达到预设数量的学校和幼托机构作为监控点。又如，选择规模（例如以日营业额统计）达到预设规模的药店作为监控点。再如，选择规模（例如以日就医人数统计）达到预设规模的医院作为监控点。

不同时间的疾病监测数据构成疾病监测的时间序列数据。例如，可以将以日为单位采集到的疾病监测数据构成疾病监测的时间序列数据。或者，可以将以周为单位采集到的疾病监测数据构成疾病监测的时间序列数据。

医疗机构（主要包括医院）是最能捕捉疾病早期暴发预兆的场所，是开展疾病监测的首选。可以根据病人就诊情况，获取疾病监测数据。

一部分病人会自行去药店购药来缓解早期症状，因此，可以根据药店的药品销售情况，获取疾病监测数据。

儿童和青少年是疾病的高危人群以及疾病传播过程中的重要环节，也应该加强对该人群的监测。学校和幼托机构是监测儿童和青少年疾病发病情况的较佳场所。可以根据学校和幼托机构的儿童和青少年的请假情况，获得疾病监测数据。

因此，本申请中主要选择医疗机构、学校和幼托机构、药店这三类场所进行疾病监测数据的采集。当然，上述对数据源的选择，并不能限制在另外的实施方案中增加或替换其他重点关注人群或场所作为监测的数据源。例如，可以将宾馆纳入疾病监测范围，获取宾馆入住人员的疾病监测数据。

根据需要，可以取任意一类监控点（例如医疗机构）采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以取医院采集的疾病监测数据构成疾病监测的时间序列数据。或者，可以结合多类监控点采集的疾病监测数据构成疾病监测的时间序列数据。例如，可以以医院采集的疾病监测数据为主，以药店参加的疾病监测数据作为补充，构成疾病监测的时间序列数据。

疾病监测数据可以包括疾病的就诊数、就诊率、发病数、发病率等患病数据。例如，可以从医疗机构（例如医院）获取疾病（例如流感）的每日就诊数，将疾病（例如流感）的每日就诊数作为疾病监测数据。又如，可以从学校获取学生的疾病（例如流感）的每日发病数，将疾病（例如流感）的每日发病数作为疾病监测数据。

需要说明的是，时间点 0 和时间点 t 表示时间序列数据的起始时间和终止时间，并非对时间序列数据时间上的限定。可以将任意时间点作为时间点 0。

计算单元 202, 用于选取时间窗大小 w , 针对时间点 w 至时间点 t 中的每个时间点 i , 计算所述时间点 i 对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 所述时间点 i 对应的的时间窗的大小为 w , $i = w, w+1, \dots, t$ 。

时间窗为给定时间点之前邻近的一段时间。例如, 设时间窗大小 w 为 4, 给定时间点为 $t=10$, 时间窗就是 $t=6$ 到 $t=9$ 这段时间; 给定时间点为 $t=11$, 时间窗就是 $t=7$ 到 $t=11$ 这段时间; 依此类推。

时间窗大小 w 是可调参数, 可以根据实际情况进行调节。

在一实施例中, 所述疾病监测数据为每日的疾病监测数据 (例如疾病的每日就诊数), 所述时间窗大小 w 可以取为 7 (即一周)。

在另一实施例中, 所述疾病监测数据为每周的疾病监测数据 (例如疾病的每周就诊数), 所述时间窗大小 w 可以取为 3 (即三周)。

计算单元 202, 还用于根据每个时间点 i 对应的的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 计算所述时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i , 得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$, 其中

$$mz_i = \frac{|x_i - \mu_i|}{\sigma_i};$$

计算单元 202 从时间点 w 开始计算各个时间点对应的的时间窗内疾病监测数据的均值和标准差, 以及从时间点 w 开始计算各个时间点对应的疾病监测数据的 moving-zscore 值。这是因为从 $t=0$ 到 $t=w-1$ 这段时间内得到的时间窗的长度小于 w , 没有足够的 w 个疾病监测数据来计算疾病监测数据的均值和标准差。

确定单元 203, 用于根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值。

可以计算所述 mz 列表的预设分位数或预设百分位数, 将所述预设分位数或预设百分位数作为所述异常值阈值。例如, 计算 mz 列表的第 0.997 分位数, 将 mz 列表的第 0.997 分位数作为异常值阈值。又如, 计算 mz 列表的第 0.5 分位数 (即中位数), 将 mz 列表的第 0.5 分位数 (即中位数) 作为异常值阈值。

或者, 可以计算 mz 列表的平均数, 将所述平均数作为异常值阈值。

判断单元 204, 用于判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值, 若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值, 则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

例如, 异常值阈值为 12, 时间点 9 对应的疾病监测数据 x_9 的 moving-zscore 值为 13, 大于所述异常值阈值, 则疾病监测数据 x_9 为异常值。

举例说明:

29 个连续时间点的疾病监测数据为：

[3,3,4,5,4,6,7,6,7,8,9,8,17,8,7,9,10,11,12,12,13,12,14,14,15,16,16,17,16]。

取时间窗大小 $w=4$ ，也就是从第 5 个数 4 开始计算 moving-zscore 值，对于第 5 个数，其时间窗内的数为[3,3,4,5]，均值为 3.75，标准差为 0.829，则 moving-zscore 值为 $(4-3.75)/0.829=0.302$ （四舍五入保留小数点后三位）。

按照同样的方法，依次计算出第 5 个数后面所有的 moving-zscore 值，得到 mz 列表：

[0.302,2.828,2.713,0.447,1.147,3.0,2.828,0.447,12.728,0.662,0.927,0.246,0.063,2.236,1.859,1.342,2.110,0.0,4.041,1.508,2.110,2.065,1.508,2.111,0.0]。

再取 mz 列表的 0.995 分位数作为异常值阈值，得到 11.685。

列表中只有 12.728 是大于这个异常值阈值的，12.728 对应的疾病监测数据 17，所以 17 就是异常值。

实施例二的疾病异常数据检测装置获取时间点 0 到时间点 t 的疾病监测数据，构成时间序列数据 X ， $X = [x_0, x_1, x_2, \dots, x_t]$ ；选取时间窗大小 w ，针对时间点 w 至时间点 t 中的每个时间点 i ，计算所述时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，所述时间点 i 对应的时间窗的大小为 w ， $i = w, w+1, \dots, t$ ；根据每个时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，计算所述时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i ，得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$ ；根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值；判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值，若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值，则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

实施例二引入了时间窗，将最具有参考价值的历史邻近数据分布来作为参考，检测结果更加合理。由于时间窗内数据量较少，导致时间窗内的数据很难满足正态分布，因而不能采用传统的 zscore 方法检测疾病监测数据中的异常值。实施例二计算多个疾病监测数据的 moving-zscore 值，保证了数量足够大，能够满足正态分布的条件，然后再根据所有 moving-zscore 值确定异常值阈值，从而检测出疾病监测数据中的异常值。因此，实施例二对于疾病监测数据的异常检测具有更高的可适用性，实现了高效准确的疾病监测数据异常检测。

在另一实施例中，可以结合其他的疾病异常数据检测装置（其他的疾病异常数据检测装置采用不同于实施例一的疾病异常数据检测方法）进行疾病异常数据检测。具体地，实施例二的疾病异常数据检测装置还可以包括：检测单元，用于获取其他的疾病异常数据检测装置对所述时间序列数据进行检测得到的异常值；比较单元，用于将实施例二的疾病异常数据检测装置得到

的异常值与所述的疾病异常数据检测装置得到的异常值进行比较，根据实施例二的疾病异常数据检测装置得到的异常值与所述的疾病异常数据检测装置得到的异常值的比较结果得到最终的异常值。

例如，采用实施例二的疾病异常数据检测装置对疾病监测的时间序列数据进行疾病异常数据检测得到第一异常值，采用其他的疾病异常数据检测装置对所述疾病监测的时间序列数据进行疾病异常数据检测得到第二异常值，比较所述第一异常值和所述到第二异常值得到最终的异常值。若所述第一异常值和所述到第二异常值一致，则以所述第一异常值/到第二异常值作为最终的异常值；否则，若所述第一异常值和所述到第二异常值不一致，则舍弃所述第一异常值与所述第二异常值。

又如，采用实施例二的疾病异常数据检测装置对疾病监测的时间序列数据进行疾病异常数据检测得到第一异常值，采用第二疾病异常数据检测装置（该第二疾病异常数据检测装置例如采用 Grubbs 检测方法）对所述疾病监测的时间序列数据进行疾病异常数据检测得到第二异常值，采用第三疾病异常数据检测装置（该第三疾病异常数据检测装置例如采用叶斯检测方法）对所述疾病监测的时间序列数据进行疾病异常数据检测得到第三异常值，比较所述第一异常值、所述第二异常值和所述第三异常值得到最终的异常值。若所述第一异常值、所述第二异常值和所述第三异常值中至少两个异常值一致，则以一致的异常值作为最终的异常值。

在实施例一的疾病异常数据检测装置的基础上，结合其他的疾病异常数据检测装置进行疾病异常数据检测，可以获得更加准确的疾病异常数据检测结果。

在另一实施例中，可以采用实施例二的疾病异常数据检测装置分别对地区和地区下属医院进行疾病异常数据检测，得到针对地区的检测结果（即地区异常值）和针对地区下属医院的检测结果（即地区下属医院异常值），比较所述地区异常值和所述地区下属医院异常值，根据所述地区异常值和所述地区下属医院异常值的比较结果得到最终的异常值。若针对地区的检测结果和针对地区下属医院的检测结果一致，则以针对地区的检测结果/针对地区下属医院的检测结果作为最终的检测结果；否则，若针对地区的检测结果和针对地区下属医院的检测结果不一致，则舍弃所述针对地区的检测结果和针对地区下属医院的检测结果。

其中，对地区进行疾病异常数据检测是根据所述地区的疾病监测的时间序列数据（例如地区的疾控部门从各个地区下属医院收集到的疾病监测数据）得到地区异常值，对地区下属医院进行疾病异常数据检测是根据所述地区下属医院的疾病监测的时间序列数据得到所述地区下属医院的异常值。

具体来说，实施例二的疾病异常数据检测装置从地区某时间段内的疾病

监测数据（例如病例数）中找出异常值，从该地区下属医院同样时间段内的疾病监测数据中找出异常值，取两个维度的交集来作为最终的检测结果。

例如，地区在 2014-3-3，2014-3-4，2014-3-5，2014-3-6 病例数分别为 160,250，170,180，该地区下属医院（例如该地区下该疾病累计就诊次数最多的医院）在这段时间的病例数分别为 130,180,125，140。从地区病例数分布来看，在 2014-3-4 这天出现了异常病例数，同时下属医院在这天也出现了异常病例数，所以综合判定在这天出现了疾病就诊异常的现象。

又如，地区在 2014-3-3，2014-3-4，2014-3-5，2014-3-6 病例数分别为 160,210，170,180，该地区下属医院（例如该地区下该疾病累计就诊次数最多的医院）在这段时间病例数分别为 130,140,125，140。从地区病例数分布来看，在 2014-3-4 这天出现了异常病例数，但是下属医院在这天并未出现异常病例数。这是由于地区下有很多家医院，地区在这一天的就诊病例数是由下属所有的医院的病例数的总和，部分医院在这一天有一个小幅度的增长，在医院维度上并不是异常，但是所有的小幅度增长后的总和在地区维度上就表现为异常了。

采用实施例二的疾病异常数据检测装置分别对地区和地区下属医院进行疾病异常数据检测，根据针对地区的检测结果和针对地区下属医院的检测结果得到最终的检测结果，可以获得更加准确的疾病异常数据检测结果。

实施例三

本实施例提供一种非易失性可读存储介质，该非易失性可读存储介质上存储有计算机可读指令，该计算机可读指令被处理器执行时实现上述疾病异常数据检测方法实施例中的步骤，例如图 1 所示的步骤 101-105：

步骤 101，获取时间点 0 到时间点 t 的疾病监测数据，构成时间序列数据 X ， $X = [x_0, x_1, x_2, \dots, x_t]$ ；

步骤 102，选取时间窗大小 w ，针对时间点 w 至时间点 t 中的每个时间点 i ，计算所述时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，所述时间点 i 对应的时间窗的大小为 w ， $i = w, w + 1, \dots, t$ ；

步骤 103，根据每个时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i ，计算所述时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i ，得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$ ；

步骤 104，根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值；

步骤 105，判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值，若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值，则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

或者，该计算机可读指令被处理器执行时实现上述装置实施例中各模块/单元的功能，例如图 2 中的单元 201-204：

获取单元 201, 用于获取时间点 0 到时间点 t 的疾病监测数据, 构成时间序列数据 X, $X = [x_0, x_1, x_2, \dots, x_t]$;

计算单元 202, 用于选取时间窗大小 w, 针对时间点 w 至时间点 t 中的每个时间点 i, 计算所述时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 所述时间点 i 对应的时间窗的大小为 w, $i = w, w + 1, \dots, t$;

所述计算单元 202, 还用于根据每个时间点 i 对应的时间窗内疾病监测数据的均值 μ_i 和标准差 σ_i , 计算所述时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值 mz_i , 得到 mz 列表 $[mz_w, mz_{w+1}, mz_{w+2}, \dots, mz_t]$, 其中

$$mz_i = \frac{|x_i - \mu_i|}{\sigma_i};$$

确定单元 203, 用于根据所述 mz 列表确定所述时间序列数据 X 的异常值阈值;

判断单元 204, 用于判断时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值是否大于所述异常值阈值, 若时间点 i 对应的疾病监测数据 x_i 的 moving-zscore 值大于所述异常值阈值, 则所述时间点 i 对应的疾病监测数据 x_i 为异常值。

实施例四

图 3 为本申请实施例四提供的计算机装置的示意图。所述计算机装置 1 包括存储器 20、处理器 30 以及存储在所述存储器 20 中并可在所述处理器 30 上运行的计算机可读指令 40, 例如疾病异常数据检测程序。所述处理器 30 执行所述计算机可读指令 40 时实现上述疾病异常数据检测方法实施例中的步骤, 例如图 1 所示的步骤 101-105。或者, 所述处理器 30 执行所述计算机可读指令 40 时实现上述装置实施例中各模块/单元的功能, 例如图 2 中的单元 201-204。

示例性的, 所述计算机可读指令 40 可以被分割成一个或多个模块/单元, 所述一个或者多个模块/单元被存储在所述存储器 20 中, 并由所述处理器 30 执行, 以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段, 该指令段用于描述所述计算机可读指令 40 在所述计算机装置 1 中的执行过程。例如, 所述计算机可读指令 40 可以被分割成图 2 中的获取单元 201、计算单元 202、确定单元 203、判断单元 204, 各单元具体功能参见实施例二。

所述计算机装置 1 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。本领域技术人员可以理解, 所述示意图 3 仅仅是计算机装置 1 的示例, 并不构成对计算机装置 1 的限定, 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, 例如所述计算机装置 1 还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 30 可以是中央处理单元(Central Processing Unit, CPU), 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。通用处理器可以是微处理器或者该处理器 30 也可以是任何常规的处理器等, 所述处理器 30 是所述计算机装置 1 的控制中心, 利用各种接口和线路连接整个计算机装置 1 的各个部分。

所述存储器 20 可用于存储所述计算机可读指令 40 和/或模块/单元, 所述处理器 30 通过运行或执行存储在所述存储器 20 内的计算机可读指令和/或模块/单元, 以及调用存储在存储器 20 内的数据, 实现所述计算机装置 1 的各种功能。所述存储器 20 可主要包括存储程序区和存储数据区, 其中, 存储程序区可存储操作系统、至少一个功能所需的应用程序 (比如声音播放功能、图像播放功能等) 等; 存储数据区可存储根据计算机装置 1 的使用所创建的数据 (比如音频数据、电话本等) 等。此外, 存储器 20 可以包括高速随机存取存储器, 还可以包括非易失性存储器, 例如硬盘、内存、插接式硬盘, 智能存储卡 (Smart Media Card, SMC), 安全数字 (Secure Digital, SD) 卡, 闪存卡 (Flash Card)、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

所述计算机装置 1 集成的模块/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, 可以存储在一个计算机可读存储介质中。基于这样的理解, 本申请实现上述实施例方法中的全部或部分流程, 也可以通过计算机可读指令来指令相关的硬件来完成, 所述的计算机可读指令可存储于一非易失性可读存储介质中, 该计算机可读指令在被处理器执行时, 可实现上述各个方法实施例的步骤。其中, 所述计算机可读指令包括计算机可读指令代码, 所述计算机可读指令代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括: 能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是, 所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减, 例如在某些司法管辖区, 根据立法和专利实践, 计算机可读介质不包括电载波信号和电信信号。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的计算机装置和方法, 可以通过其它的方式实现。例如, 以上所描述的计算机装置实施例仅仅是示意性的, 例如, 所述单元的划分, 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。

另外, 在本申请各个实施例中的各功能单元可以集成在相同处理单元中,

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相同单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请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申请。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，本申请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申请内。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。计算机装置权利要求中陈述的多个单元或计算机装置也可以由同一个单元或计算机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。

最后应说明的是，以上实施例仅用以说明本申请的技术方案而非限制，尽管参照较佳实施例对本申请进行了详细说明，本领域的普通技术人员应当理解，可以对本申请的技术方案进行修改或等同替换，而不脱离本申请技术方案的精神和范围。