

(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 7 月 16 日 (16.07.2020)



(10) 国际公布号
WO 2020/143379 A1

(51) 国际专利分类号:
G06N 3/08 (2006.01) **G06Q 40/08** (2012.01)

(21) 国际申请号: PCT/CN2019/124210

(22) 国际申请日: 2019 年 12 月 10 日 (10.12.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910015525.5 2019 年 1 月 8 日 (08.01.2019) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛
大开曼资本大厦一座四层 847 号邮箱,
Grand Cayman (KY)。

(72) 发明人: 林建滨 (LIN, Jianbin); 中国浙江省杭州市
余杭区文一西路 969 号 3 号楼 5 楼 阿里巴巴
集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL
PROPERTY LAW CORPORATION); 中国北京市
海淀区上地三街 9 号嘉华大厦 B 座 409,
Beijing 100085 (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,
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,

(54) Title: ABNORMAL DATA DETECTION METHOD AND SYSTEM

(54) 发明名称: 异常数据的检测方法及其系统

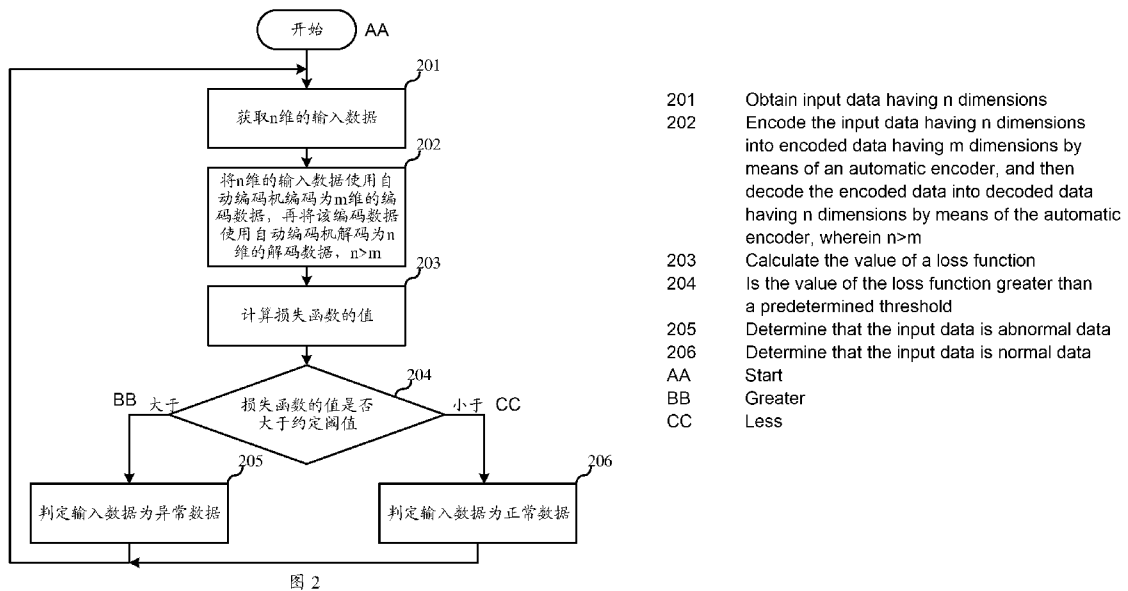


图 2

(57) Abstract: An abnormal data detection method and system, capable of effectively identifying abnormal data without labeling data and without assuming the distribution of data in advance. The method comprises: encoding input data having a first number of dimensions into encoded data having a second number of dimensions by means of an automatic encoder, and then decoding the encoded data into decoded data having the first number of dimensions by means of the automatic encoder, the first number of dimensions being greater than the second number of dimensions; calculating the value of a loss function according to the input data and the decoded data; and if the value of the loss function is greater than a predetermined threshold, determining that the input data is abnormal data.

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))。

(57) 摘要: 一种异常数据的检测方法及其系统, 无需标注数据, 也无需事先假设数据的分布, 就可以有效地识别异常数据。该方法包括: 将第一维数的输入数据使用自动编码器编码为第二维数的编码数据, 再将该编码数据使用自动编码器解码为第一维数的解码数据, 其中第一维数大于第二维数; 根据输入数据和解码数据计算损失函数的值; 如果损失函数的值大于约定阈值, 则判定输入数据为异常数据。

异常数据的检测方法及其系统

技术领域

[01] 本申请涉及计算机技术领域，特别涉及异常数据的检测技术。

背景技术

5 [02] 异常点一般指的是某个显著和其他点不同的数据点，比如图 1 中的点 A。异常点检测在很多领域都有着重要的应用，比如在保险反套利业务中，异常检测可以用来发现用户最近申请的理赔是否和大盘均值严重背离等。又如，在工业制造过程中，异常检测可以用来控制产品的质量。

[03] 目前，异常检测方法主要有两种：

10 [04] 1) 在数据中标注正常数据和异常数据，采用有监督算法学习异常检测模型。这个方法的问题在于需要大量标注数据训练模型。

[05] 2) 假设数据符合某个分布，以高斯分布为例，首先计算正常数据的均值和方差；然后计算待检测数据离正常分布的距离（即有几个标准差）。该方法的问题在于需要先假设数据符合某一个分布，然而很多数据的分布事先不可知。

15 发明内容

[06] 本申请的目的在于提供一种异常数据的检测方法及其系统，无需对大量数据打标签，也无需知道数据的分布，就可以有效地识别异常数据。

[07] 为了解决上述问题，本申请公开了一种异常数据的检测方法，包括：

20 [08] 将第一维数的输入数据使用自动编码器编码为第二维数的编码数据，再将该编码数据使用该自动编码器解码为该第一维数的解码数据，其中该第一维数大于第二维数；

[09] 根据该输入数据和该解码数据计算损失函数的值；

[10] 如果该损失函数的值大于约定阈值，则判定该输入数据为异常数据。

[11] 在一个优选例中，该损失函数是输入数据与解码数据差值的平方。

25 [12] 在一个优选例中，如果该损失函数的值小于约定阈值，则判定该输入数据为正常数据。

[13] 在一个优选例中，该输入数据是保险理赔数据，该异常数据的检测方法用于检测保险理赔是否异常。

[14] 在一个优选例中，该自动编码机的训练方法包括：

[15] 获取训练数据集合，其中包括多个该第一维数的训练数据；

5 [16] 将该训练数据集合中的每一个训练数据分别作为输入数据，将每一个训练数据分别使用该自动编码器编码为该第二维数的编码数据，再将该编码数据使用该自动编码器解码为该第一维数的解码数据；

[17] 对该自动编码机的参数进行调整，使得对于该训练数据集合中的所有训练数据和对应的解码数据，损失函数的值最小；其中，该损失函数是训练数据与对应的解码数据
10 差值的平方。

[18] 本申请还公开了一种异常数据的检测系统，包括：

[19] 自动编码器，用于将第一维数的输入数据编码为第二维数的编码数据，再将该编码数据解码为该第一维数的解码数据，其中该第一维数大于第二维数；

[20] 损失函数计算模块，用于根据该输入数据和该解码数据计算损失函数的值；

15 [21] 异常判定模块，用于在该损失函数的值大于约定阈值时，判定该输入数据为异常数据。

[22] 在一个优选例中，该损失函数是输入数据与解码数据差值的平方。

[23] 在一个优选例中，该异常判定模块还用于，在该损失函数的值小于约定阈值时，判定该输入数据为正常数据。

20 [24] 在一个优选例中，该输入数据是保险理赔数据，该异常数据的检测系统用于检测保险理赔是否异常。

[25] 在一个优选例中，还包括训练模块，用于对该自动编码器进行训练；该训练模块进一步包括：

[26] 数据获取子模块，用于获取训练数据集合，其中包括多个该第一维数的训练数据；

25 [27] 计算子模块，用于将该训练数据集合中的每一个训练数据分别作为输入数据，将每一个训练数据分别使用该自动编码器编码为该第二维数的编码数据，再将该编码数据使用该自动编码器解码为该第一维数的解码数据；

[28] 参数调整子模块，用于对该自动编码器的参数进行调整，使得对于该训练数据集合中的所有训练数据和对应的解码数据，损失函数的值最小；其中，该损失函数是训练数据与对应的解码数据差值的平方。

[29] 本申请还公开了一种异常数据的检测系统，包括：

5 [30] 存储器，用于存储计算机可执行指令；以及，

[31] 处理器，用于在执行该计算机可执行指令时实现如前文描述的方法中的步骤。

[32] 本申请还公开了一种计算机可读存储介质，该计算机可读存储介质中存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现如前文描述的方法中的步骤。

10 [33] 本申请实施方式中，无需标注数据，亦不用对数据的分布做事先假设，只要预先进行无监督训练，就可以有效地自动识别与大部分数据有明显差异的异常数据。

[34] 本申请的说明书中记载了大量的技术特征，分布在各个技术方案中，如果要罗列出本申请所有可能的技术特征的组合（即技术方案）的话，会使得说明书过于冗长。为了避免这个问题，本申请上述发明内容中公开的各个技术特征、在下文各个实施方式和例子中公开的各技术特征、以及附图中公开的各个技术特征，都可以自由地互相组合，
15 从而构成各种新的技术方案（这些技术方案均因视为在本说明书中已经记载），除非这种技术特征的组合在技术上是不可行的。例如，在一个例子中公开了特征 A+B+C，在另一个例子中公开了特征 A+B+D+E，而特征 C 和 D 是起到相同作用的等同技术手段，技术上只要择一使用即可，不可能同时采用，特征 E 技术上可以与特征 C 相组合，则，A+B+C+D 的方案因技术不可行而应当不被视为已经记载，而 A+B+C+E 的方案应当视
20 为已经被记载。

附图说明

[35] 图 1 是异常数据点的示意图。

[36] 图 2 是根据本申请第一实施方式的异常数据的检测方法流程示意图。

[37] 图 3 是根据本申请第二实施方式的异常数据的检测系统结构示意图。

25 具体实施方式

[38] 在以下的叙述中，为了使读者更好地理解本申请而提出了许多技术细节。但是，

本领域的普通技术人员可以理解，即使没有这些技术细节和基于以下各实施方式的种种变化和修改，也可以实现本申请所要求保护的技术方案。

[39] 部分概念的说明：

5 [40] 自动编码器：AutoEncoder，又可称为自编码网络，是人工神经网络的一种，用于以无监督学习的方式学习有效的数据编码。

[41] 下面概要说明本申请实施方式的部分创新点：

10 [42] 自动编码器技术原本用于图像和可视化等领域，使用该技术实现数据的降噪和降维。本申请实施方式创造性地将自动编码器应用于异常检测。自动编码器被设置为先对输入数据进行降维编码，再进行升维解码得到解码数据。使用训练数据集合训练自动编码器，使得自动编码器的损失函数最小。将待检测数据输入训练好的自动编码器，得到解码数据，进而根据该解码数据和待检测数据计算损失函数，如果该损失函数大于预定阈值，则判定为异常数据，如果该损失函数小于预定阈值，则判定为正常数据。该技术方案中自动编码器的训练是无监督训练，无需工作量巨大的数据标注工作，也不需要预先知道或假设数据所符合的分布规律，但效果十分显著。应用于保险理赔领域时可以有效地识别出异常的理赔申请。

[43] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请的实施方式作进一步地详细描述。

[44] 本申请的第一实施方式涉及一种异常数据的检测方法，其流程如图 2 所示，该方法包括以下步骤：

20 [45] 在步骤 201 中，获取输入数据 X，该输入数据的维数为第一维数（n）。

[46] 此后进入步骤 202，将 n 维的输入数据 X 使用自动编码器编码为第二维数（m）的编码数据 Z，再将该编码数据使用自动编码器解码为 n 维的解码数据 Y，其中 $n > m$ ，n 和 m 为正整数。

[47] 此后进入步骤 203，根据输入数据和解码数据计算损失函数的值。

25 [48] 损失函数的形式可以是多种多样的。在一个实施例，损失函数是输入数据与解码数据差值的平方，即 $(X-Y)^2$ 。在其他实施例中，损失函数也可以是其他的形式，例如 X 与 Y 差值的绝对值，等等。

[49] 此后进入步骤 204，判断损失函数的值是否大于约定阈值。

[50] 如果损失函数的值大于约定阈值,则进入步骤 205 判定输入数据为异常数据。

[51] 如果损失函数的值小于约定阈值,则进入步骤 206 判定输入数据为正常数据。

[52] 在一些实施例中,如果损失函数的值等于约定阈值,则判定输入数据为异常数据;在另一些实施例中,如果损失函数的值等于约定阈值,则判定输入数据为正常数据。换

5 句话说,如果损失函数的值等于约定阈值,可以根据应用场景设定其为异常数据或正常数据。

[53] 本实施方式所描述的方法无需标注数据,亦不用对数据的分布做事先假设,只要预先进行无监督训练,就可以有效地自动识别与大部分数据有明显差异的异常数据,具有显著的技术效果。

10 [54] 上面用到的自动编码器需要预先使用训练数据进行训练,训练方法如下:

[55] 获取训练数据集合,其中包括多个训练数据,每个训练数据均为 n 维。在一个实施例中,训练数据集合中的数据都是正常数据。在另一个实施例中,训练数据集合中的数据中包含少量的异常数据。这种情况符合绝大部分的应用场景,因为通常之所以称为异常数据,就是因为这部分异常数据的数据量远少于正常数据的数据量。如果异常数据

15 在训练数据集合中占比过大的话,有可能会降低本说明书一些实施例中异常数据的识别效果。

[56] 将训练数据集合中的每一个训练数据分别作为输入数据,将每一个训练数据分别使用自动编码器编码为 m 维编码数据,再将该编码数据使用自动编码器解码为 n 维解码数据。

20 [57] 对自动编码器的参数进行调整,使得对于训练数据集合中的所有训练数据和对应的解码数据,损失函数的值最小。其中,损失函数是训练数据与对应的解码数据差值的平方。

[58] 上述使用自动编码器编码和解码的步骤,以及对自动编码器参数的调整步骤,通常需要多次的迭代才可以得到使得损失函数的值最小的优化结果。

25 [59] 在一个实施例中,上述技术方案应用于保险理赔,自动编码器的输入数据是保险理赔数据(例如:最近申请理赔的次数、下单到退货的分钟数等),通过上述方法就可以检测保险理赔的申请是否异常,进而对异常的保险理赔申请进行进一步的深入调查。申请人上述技术方案应用于保险理赔的测试集数据,被判为异常的数据经过人工检测,有 42% 的样例有套现的行为,可见效果显著。

[60] 在其他一些实施例中，上述技术方案也可以应用于其他的领域，例如工业制造领域，将某种产品生产过程的各​​种监控数据作为输入数据，实时监控生产过程是否发生异常。又如，信用卡套现的检测等等。

5 [61] 本申请的第二实施方式涉及一种异常数据的检测系统，其结构如图 3 所示，该异常数据的检测系统包括：

[62] 自动编码器 301，用于将第一维数（n）的输入数据 X 编码为第二维数（m）的编码数据 Z，再将该编码数据解码为 n 维的解码数据 Y，其中 $n > m$ ，n 和 m 为正整数。

[63] 损失函数计算模块 302，用于根据输入数据和解码数据计算损失函数的值。

10 [64] 损失函数的形式可以是多种多样的。在一个实施例中，损失函数是输入数据与解码数据差值的平方，即 $(X-Y)^2$ 。在其他实施例中，损失函数也可以是其他的形式，例如 X 与 Y 差值的绝对值，等等。

[65] 异常判定模块 303，用于在损失函数的值大于约定阈值时，判定输入数据为异常数据；在损失函数的值小于约定阈值时，判定输入数据为正常数据。

15 [66] 在一些实施例中，如果损失函数的值等于约定阈值，则判定输入数据为异常数据。在另一些实施例中，如果损失函数的值等于约定阈值，则判定输入数据为正常数据。换句话说，如果损失函数的值等于约定阈值，可以根据应用场景设定其为异常数据或正常数据。

[67] 本实施方式所描述的系统无需标注数据，亦不用对数据的分布做事先假设，只要预先进行无监督训练，就可以有效地自动识别与大部分数据有明显差异的异常数据

20 [68] 图 3 中 h1、h2、h3 和 h4 代表在自动编码器（一种人工神经网络）内部各个层（例如隐藏层）中的中间结果。代表 X、Y、Z、h1、h2、h3 和 h4 的方框的高度示意了数据的维数，从图中可以形象地看出输入的 X 在自动编码器的各个层中不断地降维，一直到编码结果 Z，再从 Z 开始在自动编码器此后的各个层中不断地升维，一直到解码结果 Y。

25 [69] 在一个实施例中，上述异常数据的检测系统应用于保险理赔，自动编码器的输入数据是保险理赔数据，通过上述系统就可以检测保险理赔的申请是否异常，进而对异常的保险理赔申请进行进一步的深入调查。

[70] 在其他一些实施例中，上述系统也可以应用于其他的领域，例如工业制造领域，将某种产品生产过程的各​​种参数作为输入数据，实时监控生产过程是否发生异常。

[71] 上述系统还包括训练模块，用于对自动编码器进行训练。训练模块进一步包括：

[72] 数据获取子模块，用于获取训练数据集合，其中包括多个第一维数的训练数据。

[73] 计算子模块，用于将训练数据集合中的每一个训练数据分别作为输入数据，将每一个训练数据分别使用自动编码器编码为第二维数的编码数据，再将该编码数据使用自

5 动编码器解码为第一维数的解码数据。

[74] 参数调整子模块，用于对自动编码器的参数进行调整，使得对于训练数据集合中的所有训练数据和对应的解码数据，损失函数的值最小。其中，损失函数是训练数据与对应的解码数据差值的平方。

[75] 在一个实施例中，训练数据集合中的数据都是正常数据。

10 [76] 在另一个实施例中，训练数据集合中的数据中包含少量的异常数据。这种情况符合绝大部分的应用场景，因为通常之所以称为异常数据，就是因为这部分异常数据的数据量远少于正常数据的数据量。如果异常数据在训练数据集合中占比过大的话，有可能会降低本说明书一些实施例中

15 [77] 第一实施方式是与本实施方式相对应的方法实施方式，第一实施方式中的技术细节可以应用于本实施方式，本实施方式中的技术细节也可以应用于第一实施方式。

[78] 需要说明的是，本领域技术人员应当理解，上述异常数据的检测系统的实施方式中所示的各模块的实现功能可参照前述异常数据的检测方法的相关描述而理解。上述异常数据的检测系统的实施方式中所示的各模块的功能可通过运行于处理器上的程序（可执行指令）而实现，也可通过具体的逻辑电路而实现。本申请实施例上述异常数据的检测系统如果以软件功能模块的形式实现并作为独立的产品销售或使用

20 时，也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，本申请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）执行本申请各个实施例所述方法的全部或部分。而前述的存储介质包括：

25 U 盘、移动硬盘、只读存储器（ROM，Read Only Memory）、磁碟或者光盘等各种可以存储程序代码的介质。这样，本申请实施例不限制于任何特定的硬件和软件结合。

[79] 相应地，本申请实施方式还提供一种计算机可读存储介质，其中存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现本申请的各方法实施方式。计算机

可读存储介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括但不限于,相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读存储介质不包括暂存电脑可读媒体(transitory media),如调制的数据信号和载波。

10 [80] 此外,本申请实施方式还提供一种异常数据的检测系统,其中包括用于存储计算机可执行指令的存储器,以及,处理器;该处理器用于在执行该存储器中的计算机可执行指令时实现上述各方法实施方式中的步骤。其中,该处理器可以是中央处理单元(Central Processing Unit,简称“CPU”),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,简称“DSP”)、专用集成电路(Application Specific Integrated
15 Circuit,简称“ASIC”)等。前述的存储器可以是只读存储器(read-only memory,简称“ROM”)、随机存取存储器(random access memory,简称“RAM”)、快闪存储器(Flash)、硬盘或者固态硬盘等。本发明各实施方式所公开的方法的步骤可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。

[81] 需要说明的是,在本专利的申请文件中,诸如第一和第二等之类的关系术语仅仅
20 用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。

25 [82] 在没有更多限制的情况下,由语句“包括一个”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。本专利的申请文件中,如果提到根据某要素执行某行为,则是指至少根据该要素执行该行为的意思,其中包括了两种情况:仅根据该要素执行该行为、和根据该要素和其它要素执行该行为。多个、多次、多种等表达包括2个、2次、2种以及2个以上、2次以上、2种以上。

30 [83] 在本申请提及的所有文献都被认为是整体性地包括在本申请的公开内容中,以便

在必要时可以作为修改的依据。此外应理解，以上所述仅为本说明书的较佳实施例而已，并非用于限定本说明书的保护范围。凡在本说明书一个或多个实施例的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本说明书一个或多个实施例的保护范围之内。

- 5 [84] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描述的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

权利要求书

1. 一种异常数据的检测方法，包括：

将第一维数的输入数据使用自动编码器编码为第二维数的编码数据，再将该编码数据使用所述自动编码器解码为所述第一维数的解码数据，其中所述第一维数大于第二维
5 数；

根据所述输入数据和所述解码数据计算损失函数的值；

如果所述损失函数的值大于约定阈值，则判定所述输入数据为异常数据。

2. 如权利要求 1 所述的方法，其中，所述损失函数是输入数据与解码数据差值的平方。

10 3. 如权利要求 1 所述的方法，其中，如果所述损失函数的值小于约定阈值，则判定所述输入数据为正常数据。

4. 如权利要求 1 所述的方法，其中，所述输入数据是保险理赔数据，所述异常数据的检测方法用于检测保险理赔是否异常。

5. 如权利要求 1 所述的方法，其中，所述自动编码器的训练方法包括：

15 获取训练数据集，其中包括多个所述第一维数的训练数据；

将所述训练数据集中的每一个训练数据分别作为输入数据，将每一个训练数据分别使用所述自动编码器编码为所述第二维数的编码数据，再将该编码数据使用所述自动编码器解码为所述第一维数的解码数据；

20 对所述自动编码器的参数进行调整，使得对于所述训练数据集中的所有训练数据和对应的解码数据，损失函数的值最小；其中，所述损失函数是训练数据与对应的解码数据差值的平方。

6. 一种异常数据的检测系统，包括：

自动编码器，用于将第一维数的输入数据编码为第二维数的编码数据，再将该编码数据解码为所述第一维数的解码数据，其中所述第一维数大于第二维数；

25 损失函数计算模块，用于根据所述输入数据和所述解码数据计算损失函数的值；

异常判定模块，用于在所述损失函数的值大于约定阈值时，判定所述输入数据为异常数据。

7. 如权利要求 6 所述的系统，其中，所述损失函数是输入数据与解码数据差值的平方。

30 8. 如权利要求 6 所述的系统，其中，所述异常判定模块还用于，在所述损失函数的值小于约定阈值时，判定所述输入数据为正常数据。

9. 如权利要求 6 所述的系统, 其中, 所述输入数据是保险理赔数据, 所述异常数据的检测系统用于检测保险理赔是否异常。

10. 如权利要求 6 所述的系统, 其中, 还包括训练模块, 用于对所述自动编码器进行训练; 所述训练模块进一步包括:

5 数据获取子模块, 用于获取训练数据集合, 其中包括多个所述第一维数的训练数据;

计算子模块, 用于将所述训练数据集合中的每一个训练数据分别作为输入数据, 将每一个训练数据分别使用所述自动编码器编码为所述第二维数的编码数据, 再将该编码数据使用所述自动编码器解码为所述第一维数的解码数据;

10 参数调整子模块, 用于对所述自动编码器的参数进行调整, 使得对于所述训练数据集合中的所有训练数据和对应的解码数据, 损失函数的值最小; 其中, 所述损失函数是训练数据与对应的解码数据差值的平方。

11. 一种异常数据的检测系统, 包括:

存储器, 用于存储计算机可执行指令; 以及,

15 处理器, 用于在执行所述计算机可执行指令时实现如权利要求 1 至 5 中任意一项所述的方法中的步骤。

12. 一种计算机可读存储介质, 其中, 所述计算机可读存储介质中存储有计算机可执行指令, 所述计算机可执行指令被处理器执行时实现如权利要求 1 至 5 中任意一项所述的方法中的步骤。

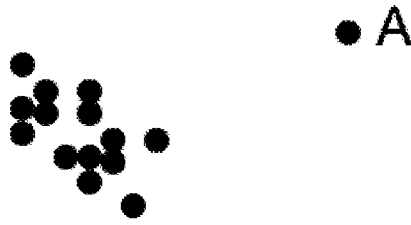


图 1

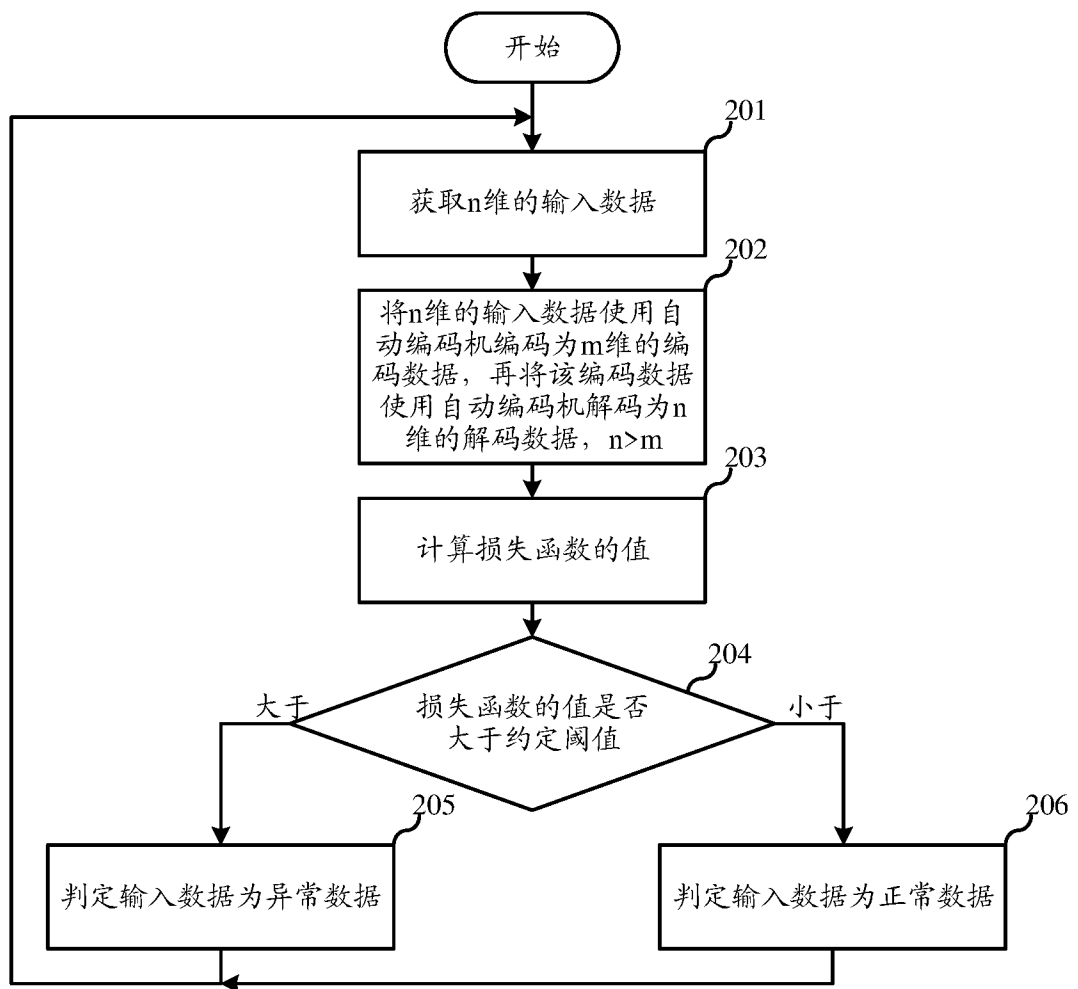


图 2

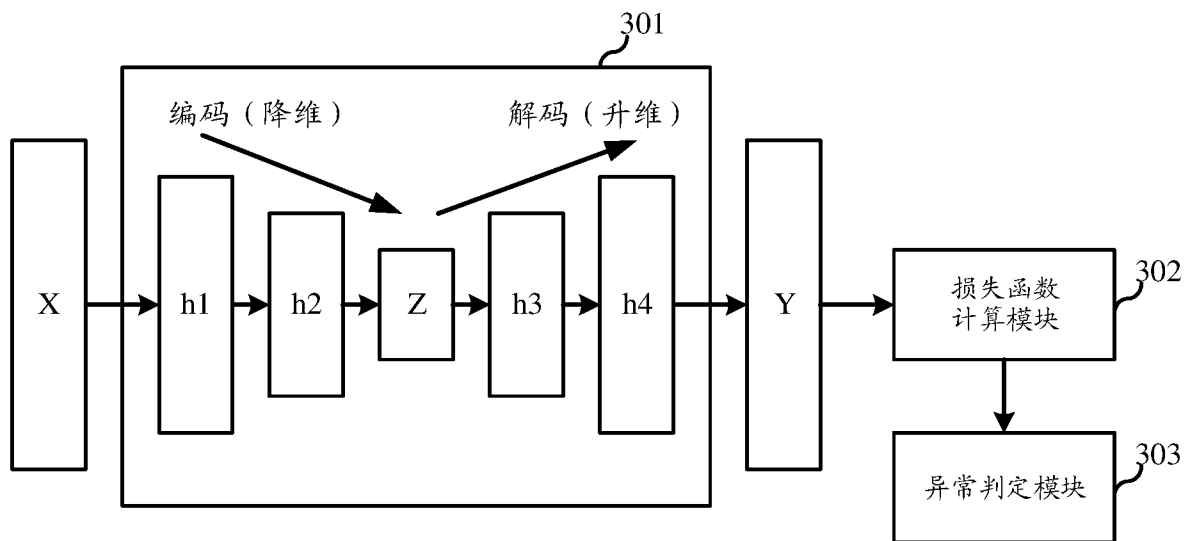


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/124210

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i; G06Q 40/08(2012.01)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)

G06N; G06Q; 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)

CNPAT, CNKI, EPODOC, WPI, IEEE: 异常, 数据, 检测, 自动, 编码, 解码, 自编码, 升维, 降维, 损失, 函数, 阈值, 平方, 训练, 参数, 保险, abnormal, data, detection, automatic, encoding, decoding, dimension, increase, reduction, loss, function, threshold, square, training, parameters, insurance

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110033014 A (ALIBABA GROUP HOLDING LIMITED) 19 July 2019 (2019-07-19) claims 1-12	1-12
X	CN 108287782 A (ZTE CORPORATION) 17 July 2018 (2018-07-17) description, paragraphs [0036]-[0079], and figures 1-5	1-12
X	CN 104008420 A (INSTITUTE OF INFORMATION ENGINEERING, CHINESE ACADEMY OF SCIENCES) 27 August 2014 (2014-08-27) description, paragraphs [0002]-[0004] and [0059]-[0103]	1-12
A	US 2018268005 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 20 September 2018 (2018-09-20) entire document	1-12



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

21 February 2020

Date of mailing of the international search report

09 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration
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/CN2019/124210

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110033014	A	19 July 2019	None			
CN	108287782	A	17 July 2018	WO	2018224055	A2	13 December 2018
CN	104008420	A	27 August 2014	None			
US	2018268005	A1	20 September 2018	CN	106776641	A	31 May 2017
				WO	2017088587	A1	01 June 2017
				EP	3373157	A1	12 September 2018

国际检索报告

国际申请号

PCT/CN2019/124210

A. 主题的分类 G06N 3/08(2006.01)i; G06Q 40/08(2012.01)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) G06N; G06Q; 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)) CNPAT, CNKI, EPDOC, WPI, IEEE: 异常, 数据, 检测, 自动, 编码, 解码, 自编码, 升维, 降维, 损失, 函数, 阈值, 平方, 训练, 参数, 保险, abnormal, data, detection, automatic, encoding, decoding, dimension, increase, reduction, loss, function, threshold, square, training, parameters, insurance																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 110033014 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求1-12</td> <td>1-12</td> </tr> <tr> <td>X</td> <td>CN 108287782 A (中兴通讯股份有限公司) 2018年 7月 17日 (2018 - 07 - 17) 说明书第[0036]-[0079]段, 附图1-5</td> <td>1-12</td> </tr> <tr> <td>X</td> <td>CN 104008420 A (中国科学院信息工程研究所) 2014年 8月 27日 (2014 - 08 - 27) 说明书第[0002]-[0004]、[0059]-[0103]段</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>US 2018268005 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2018年 9月 20日 (2018 - 09 - 20) 全文</td> <td>1-12</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 110033014 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求1-12	1-12	X	CN 108287782 A (中兴通讯股份有限公司) 2018年 7月 17日 (2018 - 07 - 17) 说明书第[0036]-[0079]段, 附图1-5	1-12	X	CN 104008420 A (中国科学院信息工程研究所) 2014年 8月 27日 (2014 - 08 - 27) 说明书第[0002]-[0004]、[0059]-[0103]段	1-12	A	US 2018268005 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2018年 9月 20日 (2018 - 09 - 20) 全文	1-12
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 110033014 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求1-12	1-12															
X	CN 108287782 A (中兴通讯股份有限公司) 2018年 7月 17日 (2018 - 07 - 17) 说明书第[0036]-[0079]段, 附图1-5	1-12															
X	CN 104008420 A (中国科学院信息工程研究所) 2014年 8月 27日 (2014 - 08 - 27) 说明书第[0002]-[0004]、[0059]-[0103]段	1-12															
A	US 2018268005 A1 (HUAWEI TECHNOLOGIES CO., LTD.) 2018年 9月 20日 (2018 - 09 - 20) 全文	1-12															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td> * 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 </td> <td> “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 </td> </tr> </table>			* 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													
* 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 21日		国际检索报告邮寄日期 2020年 3月 9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 徐淑娴 电话号码 86-(10)-53961535															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/124210

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	110033014	A	2019年 7月 19日	无			
CN	108287782	A	2018年 7月 17日	W0	2018224055	A2	2018年 12月 13日
CN	104008420	A	2014年 8月 27日	无			
US	2018268005	A1	2018年 9月 20日	CN	106776641	A	2017年 5月 31日
				W0	2017088587	A1	2017年 6月 1日
				EP	3373157	A1	2018年 9月 12日