(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112835893 B
(45)授权公告日 2023.03.21

(21)申请号 202110063078.8
(22)申请日 2021.01.18
(65)同一申请的已公布的文献号
申请公布号 CN 112835893 A
(43)申请公布日 2021.05.25
(73)专利权人 浙江大学山东工业技术研究院
地址 277000 山东省枣庄市高新区互联网小镇15号楼401房间
(72)发明人 吴健 姜晓红 应豪超 张久成
(74)专利代理机构 杭州橙知果专利代理事务所(特殊普通合伙) 33261
专利代理师 杜放
(51)Int.Cl.
G06F 16/22(2019.01)
G06F 16/2458(2019.01)
G06F 18/23213(2023.01)
G06Q 40/08(2012.01)
(56)对比文件
CN 111785384 A,2020.10.16
CN 109636653 A,2019.04.16
CN 109636644 A,2019.04.16
CN 109636192 A,2019.04.16
CN 111612636 A,2020.09.01
CN 109636650 A,2019.04.16
CN 111899114 A,2020.11.06
CN 111582879 A,2020.08.25
审查员 李国鑫

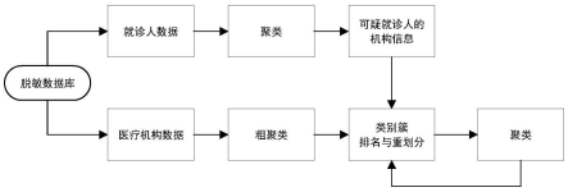
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于聚类的医保欺诈行为的检测方法及系统

(57)摘要

本发明属于医疗保险数据处理技术领域，尤其是涉及一种基于聚类的医保欺诈行为的检测方法及系统。一种基于聚类的医保欺诈行为的检测方法，包括以下步骤：S1、数据提取；S2、归一化处理；S3、数据聚类；S4、得出结果，重复步骤S3直至聚类收敛，得到可疑异常行为的机构信息。本发明提供了一种解决实际场景中推拿项目异常的问题、检测出有疑似医保欺诈异常行为的医疗机构的基于聚类的医保欺诈行为的检测方法及系统。



CN 112835893 B

1.一种基于聚类的医保欺诈行为的检测方法,其特征在于包括以下步骤:

S1、数据提取,从已脱敏数据库中提取医疗机构和就诊人的多维信息;

S2、归一化处理,对步骤S1所得的医疗机构和就诊人数据分别进行归一化处理;

S3、数据聚类,分别对步骤S2所得的医疗机构数据和就诊人数据进行聚类后,进行信息联立并训练;

所述步骤S3还包括以下步骤:

S31、对就诊人的数据进行聚类,提取疑似异常人员的相关机构信息列表;

S32、对医疗机构的数据进行粗聚类,得到疑似异常机构;

S33、联立步骤S31和步骤S32所得的疑似异常人员的相关机构信息列表和疑似异常机构信息;

S34、对步骤S32得到的疑似异常机构进行排名,将排名靠前的机构作为新的异常簇,排名靠后的机构作为新的正常簇,继续训练;

S4、得出结果,重复步骤S3直至聚类收敛,得到可疑异常行为的机构信息;

所述步骤S1中医疗机构的多维信息至少包括机构编码、机构名称、机构地址、每日的就诊项目数量、每日的就诊人数、每日的就诊人人次、每日工作时长;

所述步骤S1中就诊人的多维信息至少包括就诊人编码、就诊人姓名、年龄、疾病种类、疾病数量、项目数量、项目次数、就诊过的机构编码、就诊过的机构名称、机构数量、各机构之间的距离中位数。

2.根据权利要求1所述的一种基于聚类的医保欺诈行为的检测方法,其特征在于所述步骤S31还包括对就诊人数据聚类,得到疑似异常人员信息,进而得到其去过的机构信息列表。

3.根据权利要求1所述的一种基于聚类的医保欺诈行为的检测方法,其特征在于所述步骤S33还包括取疑似异常机构信息与机构信息列表的交集,得到交集的机构信息。

4.根据权利要求3所述的一种基于聚类的医保欺诈行为的检测方法,其特征在于所述步骤S34还包括利用步骤S33所得的交集机构信息对机构聚类的类簇重新划分,将属于机构交集的机构划分为新的异常簇,除此之外的机构划分为新的正常簇,继续训练模型。

5.一种基于聚类的医保欺诈行为的检测系统,其特征在于包括:

存储器,存储计算机可执行指令以及在执行所述计算机可执行指令时使用或生产的数据;

处理器,与所述存储器通信连接,并配置为执行存储器存储的计算机可执行指令;

所述计算机可执行指令在被执行时,实现如权利要求1～4中任一一项所述的检测方法。

一种基于聚类的医保欺诈行为的检测方法及系统

技术领域

[0001] 本发明属于医疗保险数据处理技术领域，尤其是涉及一种基于聚类的医保欺诈行为的检测方法及系统。

背景技术

[0002] 医保对解决民生问题发挥着重要作用。但伴随着医保的推进，我国基本医疗保险制度不断健全，已经建立起覆盖广泛的全民基本医疗保险体系。医保基金收入的增长的同时，受人口老龄化加护、保障水平提升等方面因素的影响，为医保基金的长期稳定运营带来较大的压力。受监管制度体系不健全，激励约束机制不完善等因素制约，医保基金使用效率不高，欺诈骗保问题频发普发，基金监管形势较为严峻，而医保欺诈是指参保人或组织机构在参加医疗保险、享受医疗保险待遇的过程中，弄虚作假、虚假开药、盗刷套刷并隐瞒真实情况等造成医疗保险基金损失的行为。

[0003] 为了提升医保基金使用效率，遏制欺诈骗保问题频发普发的现象，引入了大数据及相关的机器学习技术。它能够解决医保人员面对海量数据时数据复杂、分析难度过高、耗时等问题，利用模型和大数据分析手段，可以直接对欺诈骗保可疑案例进行筛查。医保消费中异常行为多种多样，其中推拿针灸的理疗项目中存在异常消费的情况，这类行为可能由医疗机构单方面或医疗机构与就诊人双方联合造成的，这类行为有一定的时序性、区域性，从单条记录来看无法发现异常性，此时需要使用到大数据的方法从费用、项目类型、时序等方面检测可能的异常行为，避免医保局相关人员在大量数据中查找的情况，医保局相关人员只需审核可疑机构的相关数据即可，大大减轻其工作量，提升工作效率。

发明内容

[0004] 本发明所要解决的技术问题是提供一种解决实际场景中推拿项目异常的问题、检测出有疑似医保欺诈异常行为的医疗机构的基于聚类的医保欺诈行为的检测方法及系统。为此，本发明采用以下技术方案：

[0005] 一种基于聚类的医保欺诈行为的检测方法，包括以下步骤：

[0006] S1、数据提取，从已脱敏数据库中提取医疗机构和就诊人的多维信息；

[0007] S2、归一化处理，对步骤S1所得的医疗机构和就诊人数据分别进行归一化处理；

[0008] S3、数据聚类，分别对步骤S2所得的医疗机构数据和就诊人数据进行聚类后，进行信息联立并训练；

[0009] S4、得出结果，重复步骤S3直至聚类收敛，得到可疑异常行为的机构信息。

[0010] 由于带有个人敏感信息的原始数据存储在政府医保系统中，而要对这些数据进行处理，需要在进行数据充分脱敏的前提下导出到工作系统当中，再转至安全的工作系统中进行数据的存储。

[0011] 在采用上述技术方案的基础上，本发明还可采用以下进一步的技术方案：

[0012] 所述步骤S1中医疗机构的多维信息至少包括机构编码、机构名称、机构地址、每日

的就诊项目数量、每日的就诊人数、每日的就诊人人次、每日工作时长。

[0013] 所述步骤S1中就诊人的多维信息至少包括就诊人编码、就诊人姓名、年龄、疾病种类、疾病数量、项目数量、项目次数、就诊过的机构编码、就诊过的机构名称、机构数量、各机构之间的距离中位数。

[0014] 所述步骤S2中除机构编码、机构名称、就诊人编码、就诊人名称之外，根据各个属性的差异考虑，需要对数据进行适当的归一化操作。

[0015] 所述步骤S3还包括以下步骤：

[0016] S31、对就诊人的数据进行聚类，提取疑似异常人员的相关机构信息列表；

[0017] S32、对医疗机构的数据进行粗聚类，得到疑似异常机构；

[0018] S33、联立步骤S31和步骤S32所得的疑似异常人员的相关机构信息列表和疑似异常机构信息；

[0019] S34、对步骤S32得到的疑似异常机构进行排名，将排名靠前的机构作为新的异常簇，排名靠后的机构作为新的正常簇，继续训练。

[0020] 所述步骤S31还包括对就诊人数据聚类，得到疑似异常人员信息，进而得到其去过的机构信息列表。

[0021] 所述步骤S33还包括取疑似异常机构信息与机构信息列表的交集，得到交集的机构信息。

[0022] 所述步骤S34还包括利用步骤S33所得的交集机构信息对机构聚类的类簇重新划分，将属于机构交集的机构划分为新的异常簇，除此之外的机构划分为新的正常簇，继续训练模型。

[0023] 具体地，使用k-means方法进行聚类，使用余弦距离计算两类向量之间的距离，距离较近的划分至一个类簇，最终以类簇之间的距离最大为最佳效果，其中k值为2。首先对就诊人数据聚类，得到疑似异常人员信息，进而得到其去过的机构信息列表。然后对机构数据聚类，得到疑似机构信息，取疑似异常机构信息与机构信息列表的交集，得到交集的机构信息。利用交集的机构信息对机构聚类的类簇重新划分，将属于机构交集的机构划分为新的异常簇，除此之外的机构划分为新的正常簇，用这种方式继续训练模型直至收敛。

[0024] 所述余弦距离的计算公式如下：

[0025] $$cos\theta = \frac{a \cdot b}{||a|| \times ||b||}$$

[0026] 其中，a、b是两个不同的特征向量。

[0027] 本发明还同时提供以下技术方案：

[0028] 一种基于聚类的医保欺诈行为的检测系统，包括：

[0029] 存储器，存储计算机可执行指令以及在执行所述计算机可执行指令时使用或生产的数据；

[0030] 处理器，与所述存储器通信连接，并配置为执行存储器存储的计算机可执行指令；

[0031] 所述计算机可执行指令在被执行时，实现上述的检测方法。

[0032] 与现有技术相比，本发明具有以下有益效果：

[0033] (1)联立就诊人的情况来反查机构，利用就诊人的大量数据来弥补机构数量较少的情况；

[0034] (2)提出了一种辅助聚类的方法,利用可疑就诊人的区分效果来辅助机构数据的聚类,用以提升机构聚类的可靠性;

[0035] (3)某种程度上,可快速定位可疑行为,极大地减轻了医保稽查人员的工作量,提高工作效率。

附图说明

[0036] 图1为本发明一种基于聚类的医保欺诈行为的检测方法及系统的聚类方案结构图。

具体实施方式

[0037] 为了进一步理解本发明,下面结合具体实施方式对本发明提供的一种基于聚类的医保欺诈行为的检测方法及系统进行具体描述,但本发明并不限于此,该领域技术人员在本发明核心指导思想下做出的非本质改进和调整,仍然属于本发明的保护范围。

[0038] 实施例一,一种基于聚类的医保欺诈行为的检测方法,包括以下步骤:

[0039] S1、数据提取,从已脱敏数据库中提取医疗机构和就诊人的多维信息,针对就诊人、医疗机构的数据,各自整理成按人为单位的数据、按机构为单位的数据。

[0040] 具体地,包括以下两部分:

[0041] 从脱敏数据库中抽取指定时间段内的医疗机构的多维信息,具体信息包括机构编码、机构名称、机构地址、每日的就诊项目数量、每日的就诊人数、每日的就诊人人次、机构每日工作时长。

[0042] 以及,从脱敏数据库中抽取指定时间段内的就诊人的多维信息,具体信息包括就诊人编码、就诊人姓名、年龄、疾病种类、疾病数量、项目数量、项目次数、就诊过的机构编码、就诊过的机构名称、机构数量、各机构之间的距离中位数。

[0043] 其中,需要对部分变量的含义做解释说明。假设就诊人某一个时间段内共去了4家医疗机构,则两两计算机构之间的欧式距离,然后取距离数值的中位数。

[0044] S2、归一化处理,对步骤S1所得的医疗机构和就诊人数据分别进行归一化处理。

[0045] S3、数据聚类,分别对步骤S2所得的医疗机构数据和就诊人数据进行聚类后,进行信息联立并训练。

[0046] 具体地,步骤S3包括以下步骤:

[0047] S31、对就诊人的数据进行聚类,得到疑似异常人员信息,进而得到其去过的机构信息列表。

[0048] S32、对医疗机构的数据进行粗聚类,得到疑似异常机构;

[0049] S33、联立步骤S31和步骤S32所得的疑似异常人员的相关机构信息列表和疑似异常机构信息,取疑似异常机构信息与机构信息列表的交集,得到交集的机构信息。

[0050] S34、对步骤S32得到的疑似异常机构进行排名,将排名靠前的机构作为新的异常簇,排名靠后的机构作为新的正常簇,即利用步骤S33所得的交集机构信息对机构聚类的类簇重新划分,将属于机构交集的机构划分为新的异常簇,除此之外的机构划分为新的正常簇,继续训练模型。

[0051] 具体地,使用k-means方法进行聚类,使用余弦距离计算两类向量之间的距离,距

离较近的划分至一个类簇，最终以类簇之间的距离最大为最佳效果，其中k值为2。首先先对就诊人的数据进行聚类处理，待收敛之后提取可疑人员去过的医疗机构信息，得到医疗机构信息列表；然后对医疗机构的数据进行粗聚类，得到初步的正常与异常机构信息，对正常与异常机构进行排名，异常机构排名靠前、正常机构排名靠后。联立医疗机构信息列表，若机构出现在医疗机构信息列表且此机构原本在异常类型中，则排名不变；若机构不在医疗信息列表且此机构原本在异常类型中，则将此机构移动至异常机构排名末端；若机构在医疗机构信息列表且此机构原本在正常类型中，则将此机构移动至异常机构排名末端；然后分别取异常排名前端的5家机构、正常排名前端的5家机构各自组成初始中心点，继续聚类训练直至收敛。

[0052] K-Means(K均值)聚类算法具体步骤介绍如下：

[0053] Step 1:确定聚类类别为2，并随机初始化2个中心点；

[0054] Step 2:计算每个数据点到2个中心点的距离，将每个数据点分配给距离该数据点较近的中心类中；

[0055] Step 3:重新计算2类各自的中心点；

[0056] Step 4:重复以上Step 2→Step 3，直到每一类的中心点在每次迭代后变化不大；

[0057] 所述余弦距离的计算公式如下：

[0058] $$cos\theta = \frac{a \cdot b}{||a|| \times ||b||}$$

[0059] 其中，a、b是两个不同的特征向量。

[0060] S4、得出结果，重复步骤S3直至聚类收敛，得到可疑异常行为的机构信息。

[0061] 实施例二，一种基于聚类的医保欺诈异常行为的检测系统，包括：

[0062] 存储器，存储计算机可执行指令以及在执行所述计算机可执行指令时使用或生产的数据；

[0063] 处理器，与所述存储器通信连接，并配置为执行存储器存储的计算机可执行指令，

[0064] 所述计算机可执行指令在被执行时，实现实施例一的基于聚类的医保欺诈异常行为的检测方法。

[0065] 虽然本发明已通过参考优选的实施例进行了图示和描述，但是，本专业普通技术人员应当了解，在权利要求书的范围内，可作形式和细节上的各种各样变化。

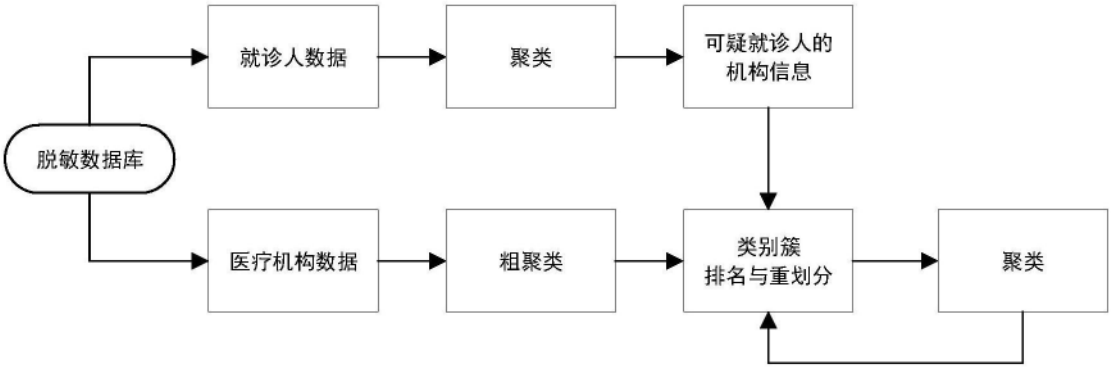


图1