(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116894271 A
(43)申请公布日 2023.10.17

(21)申请号 202310980143.2

(22)申请日 2023.08.04

(71)申请人 中国医学科学院医学信息研究所
地址 100020 北京市朝阳区雅宝路3号

(72)发明人 吴思竹 唐明坤 钱庆

(74)专利代理机构 北京睿智保诚专利代理事务所(普通合伙) 11732
专利代理师 韩迎之

(51)Int.Cl.
G06F 21/62(2013.01)
G06F 18/23(2023.01)

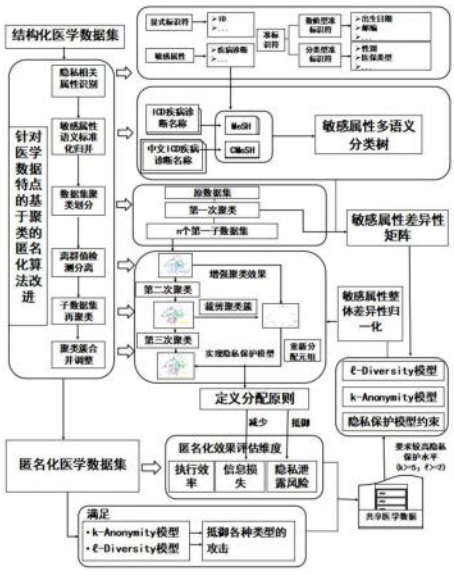
权利要求书3页 说明书16页 附图5页

(54)发明名称

一种基于匿名化算法的数据共享隐私保护方法

(57)摘要

本发明公开了一种基于匿名化算法的数据共享隐私保护方法,涉及隐私保护技术领域,包括:识别和归并需要进行隐私保护的属性,明确匿名化处理对象;采用MSAK匿名算法第一次聚类将原始数据集划分为多个第一子数据集,分离每个第一子数据集的离群值,形成第二子数据集和第一待分配元组集合;第二次聚类对每个第二子数据集聚类,生成多个由相似元组构成的聚类簇,判断聚类簇的大小和过远元组剪裁,形成第一聚类簇集合、第二聚类簇集合和第二待分配元组集合;在所有聚类簇的基础上进行第三次聚类,将所有待分配元组分配到第一聚类簇集合、第二聚类簇集合中,用聚类中心取代聚类簇内所有元组,生成满足隐私保护模型的等价类,从而实现数据的匿名化。



CN 116894271 A

1.一种基于匿名化算法的数据共享隐私保护方法,其特征在于,包括:

对结构化医学数据集进行隐私相关属性识别;

采用MSAK匿名算法对识别后的属性进行语义标准化归并,构建敏感属性多语义分类树,计算敏感属性最小差异性;

构建虚拟初始聚类中心,对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理,得到第一子数据集和敏感属性差异性矩阵;

对所述第一子数据集进行离群值检测分离处理,得到第二子数据集以及由离群值构成的第一待分配元组集合;

设定k-Anonymity模型的参数k和l-Diversity模型的参数l,对所述第二子数据集进行聚类,将所述第二子数据集的元组聚集成聚类簇,根据聚类簇的大小、参数k和参数l对所述聚类簇进行裁剪和判断,得到第一聚类簇集合、第二聚类簇集合和第二待分配元组集合;

根据聚类簇的元组数量大小判断是否满足k-Anonymity模型;根据所述敏感属性差异性矩阵,计算聚类簇敏感属性值的整体差异性,判断聚类簇中敏感属性值的整体差异性是否满足l-Diversity模型;

将所述待分配元组集合中的元组重新逐个分配到第一聚类簇集合和第二聚类簇集合中,并对第一聚类簇集合中的剩余聚类簇进行合并调整,经过泛化和抑制后,得到匿名化的数据集。

2.根据权利要求1所述的一种基于匿名化算法的数据共享隐私保护方法,其特征在于,对结构化医学数据集进行隐私相关属性识别,包括:确定结构化医学数据集中需要进行匿名化的所有属性,将其中的显式标识符、准标识符和/或敏感属性识别出来,然后再按照这些类型分别对属性进行处理;

其中,所述准标识符包括数值型准标识符和分类型准标识符;

所述数值型准标识符的距离度量采用欧几里得距离、曼哈顿距离或切比雪夫距离的计算方法得到;

所述分类型准标识符的距离度量基于相应的泛化层次结构树确定。

3.根据权利要求2所述的一种基于匿名化算法的数据共享隐私保护方法,其特征在于,所述构建虚拟初始聚类中心,对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理,包括:统计每个准标识符的取值大小或各属性值出现的频率;将所述数值型准标识符依据数值从小到大升序排列,将所述分类型准标识符依据各属性值出现频率从小到大按比例升序排列,组建序列;

设置需要划分的第一子数据集数量为n,对各序列等间距选取n个值;

每部分的中线对应的各序列的值即为虚拟初始聚类中心;

将序列中的所有元组逐个与聚类中心进行距离比较,将每个元组纳入距离最近的聚类簇中,并更新该聚类簇的聚类中心;

记录得到非重复敏感属性值,计算每两个敏感属性值的差异性,构建敏感属性差异性矩阵。

4.根据权利要求2所述的一种基于匿名化算法的数据共享隐私保护方法,其特征在于,对所述第一子数据集进行离群值检测分离处理的过程包括:

根据所述泛化层次结构树将所述第一子数据集的分类属性值转化为哑变量值;

通过孤立森林算法拟合数据，检测第一子数据集的离群值，设定离群值比例参数为o；

生成离群值集合，并将其纳入到待分配元组集合中，生成第一待分配元组集合；

从第一子数据集中分离出所述离群值集合中的数据，生成第二子数据集。

5.根据权利要求1所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，对所述第二子数据集进行聚类的过程包括：

在所述第二子数据集中，随机选取f个元组作为聚类中心；

将剩余的元组与所有聚类中心进行距离比较，并纳入到最近的聚类簇中，更新该聚类簇的聚类中心；

对于每个聚类簇，如果聚类簇的大小小于参数k，则将所述聚类簇纳入到第一聚类簇集合中；

如果聚类簇的大小大于参数k，则将距离远的元组分离出来，保留k个元组，判断k个元组是否满足l-Diversity模型，如满足则将k个元组纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中；将分离出来的距离远的元组纳入到第二待分配元组集合中；

如果聚类簇的大小等于参数k，判断该聚类簇是否满足l-Diversity模型，如满足则将该聚类簇纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中；

其中，$f=\frac{g}{k}$；g表示子数据集的元组数。

6.根据权利要求1所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，计算聚类簇敏感属性值的整体差异性的过程包括：

获取所述敏感属性差异性矩阵；

设定聚类簇中有h个敏感属性值，计算各敏感属性值相互间的差异性之和，得到聚类簇敏感属性值的整体差异性；

并对所述聚类簇敏感属性值的整体差异性进行归一化处理。

7.根据权利要求6所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，所述聚类簇敏感属性值的整体差异性，用公式表示为：

$$d=\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h}l_i l_j;$$

对所述聚类簇敏感属性值的整体差异性进行归一化处理，用公式表示为：$D=\frac{d}{h(h-1)}$；

其中，在计算整体差异性时，聚类簇中每个元素都需要比较h-1次，因此只需要保证在这h-1次中，有l-1次的差异性结果为1，便能满足l-Diversity模型；用公式表达为：

$$D_{min}=\frac{\ell-1}{h-1};$$

当整体差异性D达到D_{min}及以上时，则认为聚类簇中敏感属性值的整体差异性满足l-Diversity模型。

8.根据权利要求1所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，基于最小簇长约束原则、满足差异性约束原则以及最小信息损失原则，将所述待分配元组集合中的元组逐个分配到第一聚类簇集合和第二聚类簇集合中；

所述最小簇长约束原则是指当所述第一聚类簇集合中聚类簇的数量大于0时，先将所述待分配元组集合中的元组分配到所述第一聚类簇集合中，确保每个所述第一聚类簇集合

中聚类簇的最小簇长达到k，满足k-Anonymity模型；才能从第一聚类簇集合中剔除并纳入到所述第二聚类簇集合中；

所述满足差异性约束原则是指所述第一聚类簇集合中的聚类簇纳入所述待分配元组集合后满足l-Diversity模型，才能从第一聚类簇集合中剔除并纳入到所述第二聚类簇集合中；

所述最小信息损失原则是指将所述待分配元组集合与聚类簇的聚类中心比较距离，并分配到距离最近的聚类簇中；对剩余的第一聚类簇集合中的聚类簇进行就近合并，直至同时满足k-Anonymity模型和l-Diversity模型后纳入第二聚类簇集合；否则对聚类簇进行抑制处理。

9.根据权利要求8所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，用聚类中心取代所述第二聚类簇集合中聚类簇的所有元组，使每个聚类簇分别生成一个等价类，每个等价类由多条相同元组构成；每个等价类的大小与相应的聚类簇的大小相同；所述等价类共同构成匿名化的数据集。

10.根据权利要求1所述的一种基于匿名化算法的数据共享隐私保护方法，其特征在于，所述待分配元组集合包括所述第一待分配元组集合和所述第二待分配元组集合。

一种基于匿名化算法的数据共享隐私保护方法

技术领域

[0001] 本发明涉及隐私保护技术领域，更具体的说是涉及一种基于匿名化算法的数据共享隐私保护方法。

背景技术

[0002] 目前，随着大数据和医疗信息化建设的发展，数据共享成为了大数据利用和学术研究过程中的重要环节。医学数据涉及许多个人的生命健康安全相关信息，如何在医学数据共享过程中，实现有效的隐私保护是一个值得研究和探索的问题。近年来，各国人员都在不断加强对医学数据共享隐私保护的研究，包括数据收集、数据保存和数据使用等环节的隐私保护问题。在数据收集阶段就应该完成对医学数据的匿名化处理，保证匿名化后的数据不能复原且不能被重新识别或关联是数据隐私保护的共同要求。

[0003] 然而，匿名化处理往往会造成较大的数据质量下降。因为医学数据具有安全性、准确性、海量性、异质性和复杂性等特点，导致现有的匿名化算法在医学数据的匿名化处理过程表现较差，造成信息损失较多。

[0004] 因此，如何提供一种基于匿名化算法的数据共享隐私保护方法，尽可能地在满足隐私保护要求的前提下，减少信息损失，保留更多的数据质量是本领域技术人员亟需解决的问题。

发明内容

[0005] 有鉴于此，本发明提供了一种基于匿名化算法的数据共享隐私保护方法，针对医学数据存在的数据规模较大、离群值较多及包含多语义的疾病诊断属性等特点，本发明改进了传统基于聚类的匿名化算法流程，设计了MSAK匿名算法(多语义敏感属性K匿名算法，Multi-semantic Sensitive Attributes K-Anonymity Algorithm)，在影响匿名化算法表现的关键因素结果确定的基础上，以满足隐私保护要求的同时尽可能减少信息损失为目标。

[0006] 为了实现上述目的，本发明采用如下技术方案：一种基于匿名化算法的数据共享隐私保护方法，包括：

[0007] 对结构化医学数据集进行隐私相关属性识别；

[0008] 采用MSAK匿名算法对识别后的属性进行语义标准化归并，构建敏感属性多语义分类树，计算敏感属性最小差异性；

[0009] 构建虚拟初始聚类中心，对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理，得到第一子数据集和敏感属性差异性矩阵；

[0010] 对所述第一子数据集进行离群值检测分离处理，得到第二子数据集以及由离群值构成的第一待分配元组集合；

[0011] 设定k-Anonymity模型的参数k和ℓ-Diversity模型的参数ℓ，对所述第二子数据集进行聚类，将所述第二子数据集的元组聚集成聚类簇，根据聚类簇的大小、参数k和参数ℓ对

所述聚类簇进行裁剪和判断，得到第一聚类簇集合、第二聚类簇集合和第二待分配元组集合；所述第二待分配元组集合是在第一待分配元组集合的基础上进行更新；

[0012] 根据聚类簇的元组数量大小判断是否满足k-Anonymity模型；根据所述敏感属性差异性矩阵，计算聚类簇敏感属性值的整体差异性，判断聚类簇中敏感属性值的整体差异性是否满足ℓ-Diversity模型；

[0013] 将所述待分配元组集合中的元组重新逐个分配到第一聚类簇集合和第二聚类簇集合中，并对第一聚类簇集合中的剩余聚类簇进行合并调整，经过泛化和抑制后，得到匿名化的数据集。

[0014] 优选的，对结构化医学数据集进行隐私相关属性识别，包括：确定结构化医学数据集中需要进行匿名化的所有属性，将其中的显式标识符、准标识符和/或敏感属性识别出来，然后再按照这些类型分别对属性进行处理；

[0015] 其中，所述准标识符包括数值型准标识符和分类型准标识符；

[0016] 所述数值型准标识符的距离度量采用欧几里得距离、曼哈顿距离或切比雪夫距离的计算方法得到；

[0017] 所述分类型准标识符的距离度量基于相应的泛化层次结构树确定。

[0018] 优选的，所述构建虚拟初始聚类中心，对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理，包括：统计每个准标识符的取值大小或各属性值出现的频率；将所述数值型准标识符依据数值从小到大升序排列，将所述分类型准标识符依据各属性值出现频率从小到大按比例升序排列，组建序列；

[0019] 设置需要划分的第一子数据集数量为n，对各序列等间距选取n个值；

[0020] 每部分的中线对应的各序列的值即为虚拟初始聚类中心；

[0021] 将序列中的所有元组逐个与聚类中心进行距离比较，将每个元组纳入距离最近的聚类簇中，并更新该聚类簇的聚类中心；

[0022] 记录得到非重复敏感属性值，计算每两个敏感属性值的差异性，构建敏感属性差异性矩阵。

[0023] 数据集聚类划分过程通过第一次聚类，将原始数据集划分成n个大小相近的子数据集。

[0024] 优选的，对所述第一子数据集进行离群值检测分离处理的过程包括：

[0025] 根据所述泛化层次结构树将所述第一子数据集的分类属性值转化为哑变量值；

[0026] 通过孤立森林算法拟合数据，检测第一子数据集的离群值，设定离群值比例参数为o；

[0027] 生成离群值集合，并将其纳入到待分配元组集合中，生成第一待分配元组集合；

[0028] 从第一子数据集中分离出所述离群值集合中的数据，生成第二子数据集。

[0029] 离群值检测分离过程在数据集聚类划分过程生成的所有第一子数据集中进行。因为孤立森林算法时间复杂度较低，可以减少离群值检测分离过程对匿名化算法执行效率的影响，所以MSAK匿名算法通过孤立森林算法检测这些子数据集的离群值并将其分离。

[0030] 其中，离群值检测分离过程中各第一子数据集中独立进行，互不干扰。可以通过对各第一子数据集进行并行处理，提高算法执行效率。

[0031] 同时，离群值比例参数的确定还需要结合结构化医学数据集准标识符的特点来确

定，当结构化医学数据集中存在较多的异质性较大的准标识符时，可以适当提高离群值比例参数；当结构化医学数据集中存在较多的时间类等异质性较小的准标识符时，可以适当减小离群值比例参数。

[0032] 优选的，对所述第二子数据集进行聚类的过程包括：

[0033] 在所述第二子数据集中，随机选取f个元组作为聚类中心；

[0034] 将剩余的元组与所有聚类中心进行距离比较，并纳入到最近的聚类簇中，更新该聚类簇的聚类中心；

[0035] 对于每个聚类簇，如果聚类簇的大小小于参数k，则将所述聚类簇纳入到第一聚类簇集合中；

[0036] 如果聚类簇的大小大于参数k，则将距离远的元组分离出来，保留k个元组，判断k个元组是否满足ℓ-Diversity模型，如满足则将k个元组纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中；将分离出来的距离远的元组纳入到第二待分配元组集合中；

[0037] 如果聚类簇的大小等于参数k，判断该聚类簇是否满足ℓ-Diversity模型，如满足则将该聚类簇纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中；

[0038] 其中，$f=\frac{g}{k}$；g表示子数据集的元组数。

[0039] 对第二子数据集进行聚类的过程是MSAK匿名算法的第二次聚类过程，该过程不仅将去除了离群值后的子数据集的元组聚集成簇，而且还对这些聚类簇进行了裁剪和判断，并标记满足ℓ-Diversity模型的情况。

[0040] 优选的，计算聚类簇敏感属性值的整体差异性的过程包括：

[0041] 获取所述敏感属性差异性矩阵；

[0042] 设定聚类簇中有h个敏感属性值，计算各敏感属性值相互间的差异性之和，得到聚类簇敏感属性值的整体差异性；

[0043] 并对所述聚类簇敏感属性值的整体差异性进行归一化处理。

[0044] 优选的，所述聚类簇敏感属性值的整体差异性，用公式表示为：

[0045] $$d=\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h}l_i l_j;$$

[0046] 对所述聚类簇敏感属性值的整体差异性进行归一化处理，用公式表示为：

[0047] $$D=\frac{d}{h(h-1)};$$

[0048] 其中，在计算整体差异性时，聚类簇中每个元素都需要比较h-1次，因此只需要保证在这h-1次中，有ℓ-1次的差异性结果为1，便能满足ℓ-Diversity模型；用公式表达为：

[0049] $$D_{min}=\frac{\ell-1}{h-1};$$

[0050] 当整体差异性D达到D_{min}及以上时，则认为聚类簇中敏感属性值的整体差异性满足ℓ-Diversity模型。

[0051] 优选的，基于最小簇长约束原则、满足差异性约束原则以及最小信息损失原则，将所述待分配元组集合中的元组逐个分配到第一聚类簇集合和第二聚类簇集合中；

[0052] 所述最小簇长约束原则是指当所述第一聚类簇集合中聚类簇的数量大于0时，先将所述待分配元组集合中的元组分配到所述第一聚类簇集合中，确保每个所述第一聚类簇

集合中聚类簇的最小簇长达到k，满足k-Anonymity模型，才能从第一聚类簇集合中剔除并纳入到所述第二聚类簇集合中；

[0053] 所述满足差异性约束原则是指所述第一聚类簇集合中的聚类簇纳入所述待分配元组集合后满足ℓ-Diversity模型，才能从第一聚类簇集合中剔除并纳入到所述第二聚类簇集合中；确保所有第二聚类簇集合中聚类簇最终都能满足ℓ-Diversity模型；

[0054] 所述最小信息损失原则是指将所述待分配元组集合与聚类簇的聚类中心比较距离，并分配到距离最近的聚类簇中；对剩余的第一聚类簇集合中的聚类簇进行就近合并，直至同时满足k-Anonymity模型和ℓ-Diversity模型后纳入第二聚类簇集合；否则对聚类簇进行抑制处理。

[0055] 优选的，用聚类中心取代所述第二聚类簇集合中聚类簇的所有元组，使每个聚类簇分别生成一个等价类，每个等价类由多条相同元组构成；每个等价类的大小与相应的聚类簇的大小相同；所述等价类共同构成匿名化的数据集。

[0056] 经由上述的技术方案可知，与现有技术相比，本发明公开提供了一种基于匿名化算法的数据共享隐私保护方法，包括：对结构化医学数据集进行隐私相关属性识别；采用MSAK匿名算法对识别后的属性进行语义标准化归并，构建敏感属性多语义分类树，计算敏感属性最小差异性；构建虚拟初始聚类中心，对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理，得到第一子数据集和敏感属性差异性矩阵；对所述第一子数据集进行离群值检测分离处理，得到第二子数据集以及由离群值构成的第一待分配元组集合；设定k-Anonymity模型的参数k和ℓ-Diversity模型的参数ℓ，对所述第二子数据集进行聚类，将所述第二子数据集的元组聚集成聚类簇，根据聚类簇的大小、参数k和参数ℓ对所述聚类簇进行裁剪和判断，得到第一聚类簇集合、第二聚类簇集合和第二待分配元组集合；根据聚类簇的元组数量大小判断是否满足k-Anonymity模型；根据所述敏感属性差异性矩阵，计算聚类簇敏感属性值的整体差异性，判断聚类簇中敏感属性值的整体差异性是否满足ℓ-Diversity模型；将所述待分配元组集合中的元组重新逐个分配到第一聚类簇集合和第二聚类簇集合中，并对第一聚类簇集合中的剩余聚类簇进行合并调整，经过泛化和抑制后，得到匿名化的数据集。

[0057] 本发明具有以下有益效果：

[0058] 本发明针对医学数据存在的数据规模较大、离群值较多及包含多语义的疾病诊断属性等特点，基于聚类算法的原理设计了MSAK匿名算法。在影响匿名化算法表现的关键因素分析结果的基础上，以满足隐私保护要求的同时尽可能减少信息损失为目标，MSAK匿名算法重点针对以下几个问题进行改进：1)针对现有基于聚类的匿名化算法缺乏考虑医学数据疾病诊断属性的多语义特点导致的相似性攻击风险较高的问题，构建多语义分类树，计算多语义敏感属性的最小差异性用于ℓ-Diversity模型的判断，从而降低相似性攻击的风险；2)针对基于聚类的匿名化算法在大规模数据中执行效率较低的问题，提出控制子数据集大小的数据集划分方法，使后续聚类过程能够实现并行高效计算，提高算法匿名化处理性能；3)针对离群值较多导致的基于聚类的匿名化算法的聚类效果较差的问题，本发明基于离群值检测算法优化聚类过程，采取先分离再聚类后分配的策略，减少离群值引起的匿名化过程的信息损失。

附图说明

[0059] 为了更清楚地说明本发明实施例或现有技术中的技术方案，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，还可以根据提供的附图获得其他的附图。

[0060] 图1为本发明实施例提供的一个分类型准标识符“Marital Status”(婚姻状况)的泛化层次结构树。

[0061] 图2为本发明实施例提供的“新型冠状病毒感染”多语义分类树示例图。

[0062] 图3为本发明实施例提供的构建虚拟初始聚类中心的流程示例图。

[0063] 图4为本发明实施例提供的敏感属性差异性矩阵示例图。

[0064] 图5为本发明实施例提供的孤立森林算法检测分离离群值的过程示意图。

[0065] 图6为本发明实施例提供的二次聚类后聚类簇的处理流程图。

[0066] 图7为本发明实施例提供的待分配元组的分配流程图。

[0067] 图8为本发明实施例提供的处理前包含隐私信息的数据示意图。

[0068] 图9为本发明实施例提供的处理后的数据隐私信息被泛化并可用于数据分析的示意图。

[0069] 图10为本发明实施例提供的具体算法实现流程图。

具体实施方式

[0070] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

[0071] 相比于其他类型匿名化算法，基于聚类的匿名化算法能够单元格级别地对医学数据进行匿名化处理，减少过度泛化带来的信息损失，从而保留更多的数据质量。但在现有基于聚类的匿名化算法中，缺乏对医学数据特点的考虑。本发明主要在医学数据特点的基础上，面向医学数据共享隐私保护的需求，提出了一种能够在满足相同隐私保护模型的前提下，保留更多数据质量的基于聚类的匿名化算法，即多语义敏感属性K匿名算法(Multi-semantic Sensitive Attributes K-Anonymity Algorithm,MSAK匿名算法)。

[0072] 本发明实施例公开了一种基于匿名化算法的数据共享隐私保护方法，如图10所示，包括：

[0073] 对结构化医学数据集进行隐私相关属性识别；

[0074] 采用MSAK匿名算法对识别后的属性进行语义标准化归并，构建敏感属性多语义分类树，计算敏感属性最小差异性；

[0075] 构建虚拟初始聚类中心，对识别后的结构化医学数据集以及敏感属性多语义分类树进行数据集聚类划分处理，得到第一子数据集和敏感属性差异性矩阵；

[0076] 对所述第一子数据集进行离群值检测分离处理，得到第二子数据集以及由离群值构成的第一待分配元组集合；

[0077] 设定k-Anonymity模型的参数k和ℓ-Diversity模型的参数ℓ,对所述第二子数据集

进行聚类，将所述第二子数据集的元组聚集成聚类簇，根据聚类簇的大小、参数k和参数ℓ对所述聚类簇进行裁剪和判断，得到第一聚类簇集合、第二聚类簇集合和第二待分配元组集合；所述第二待分配元组集合是在第一待分配元组集合的基础上进行更新；

[0078] 根据聚类簇的元组数量大小判断是否满足k-Anonymity模型；根据所述敏感属性差异性矩阵，计算聚类簇敏感属性值的整体差异性，判断聚类簇中敏感属性值的整体差异性是否满足ℓ-Diversity模型；

[0079] 将所述待分配元组集合中的元组重新逐个分配到第一聚类簇集合和第二聚类簇集合中，并对第一聚类簇集合中的剩余聚类簇进行合并调整，经过泛化和抑制后，得到匿名化的数据集。

[0080] 在本发明的一个具体实施例中，所述第一聚类簇集合可以表示为未满足k-Anonymity模型和ℓ-Diversity模型的集合，所述第二聚类簇集合可以表示为已满足k-Anonymity模型和ℓ-Diversity模型的集合。

[0081] 其中，所述泛化包括数值型属性泛化和分类型属性泛化；所述抑制包括显式标识符抑制和分类型属性抑制。

[0082] 在本发明的一个具体实施例中，MSAK匿名算法的实现包括隐私相关属性识别、敏感属性语义标准化归并、数据集聚类划分、离群值检测分离、子数据集再聚类和聚类簇合并调整6个过程。隐私相关属性识别、敏感属性语义标准化归并是识别和归并需要进行隐私保护的属性的过程，主要目的是明确匿名化处理的对象，为后续的聚类距离度量、敏感属性差异性度量做准备。数据集聚类划分、离群值检测分离、子数据集再聚类和聚类簇合并调整是MSAK匿名算法的核心过程，通过离群值分离和三次聚类的方式对传统基于聚类的匿名化算法进行改进。在核心过程中，MSAK匿名算法第一次聚类(数据集聚类划分过程)将原始数据集划分为多个第一子数据集，然后分离每个第一子数据集的离群值。之后，再对每个第一子数据集的剩余元组(即第二子数据集)进行第二次聚类(子数据集再聚类过程)。第二次聚类结果生成多个由相似元组构成的聚类簇，然后需要将较大的聚类簇内距离较远的元组进行裁剪。最后在所有聚类簇的基础上进行第三次聚类(聚类簇合并调整过程)，即在所有子数据集中的聚类簇合并后，将离群值和剪裁的元组重新分配到距离最近的聚类簇中。用聚类中心取代聚类簇内所有元组之后，就能够生成多个满足隐私保护模型的等价类，从而实现了医学数据的匿名化。

[0083] 具体的，隐私相关属性识别过程：隐私相关属性识别是对结构化医学数据集中的所有属性进行匿名化需求分析，确定结构化医学数据集中需要进行匿名化的所有属性，将其中的显式标识符、准标识符和/或敏感属性识别出来，然后再按照这些类型分别对属性进行处理。

[0084] (1)显式标识符

[0085] 医学数据中存在大量的显式标识符，包括姓名类、编号类和具体联系方式类属性。因为这些显式标识符只要单个存在就能直接识别个体身份，所以在隐私相关属性识别过程中，一旦将属性判定为显式标识符后，就需要将该属性的所有值都进行隐匿处理。

[0086] (2)准标识符

[0087] 准标识符是在一定的背景知识下，能够通过组合确定个体身份的信息，是匿名化算法重点处理的对象。准标识符的距离度量决定了元组在聚类过程的距离比较结果，因此

选择准标识符距离度量的方法十分关键。根据准标识符值的数据类型，MSAK匿名算法将准标识符进一步划分为数值型准标识符和分类型准标识符，两类准标识符采用不同的距离度量方式。

[0088] ①数值型准标识符距离度量

[0089] 数值型准标识符的距离度量可以选择采用欧几里得距离、曼哈顿距离或切比雪夫距离的计算方法。欧几里得距离适用于度量连续性变量的距离，它的计算原理是对数值型准标识符的差异求平方和，然后对该平方和求平方根从而得到它们的距离。用欧几里得距离计算元组$i=(x_{i1},x_{i2},\cdots,x_{in})$和$j=(x_{j1},x_{j2},\cdots,x_{j2})$的距离公式可表示为：

[0090] $$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{in}-x_{j2})^2}$$ 公式1-1；

[0091] 曼哈顿距离是度量两个数值型准标识符元组距离的另一种方法，适合用于度量离散变量的距离。它的计算原理是用元组在各个维度上的差异绝对值之和作为准标识符的距离。用曼哈顿距离计算元组$i=(x_{i1},x_{i2},\cdots,x_{in})$和$j=(x_{j1},x_{j2},\cdots,x_{j2})$的距离公式可表示为：

[0092] $d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{in}-x_{jn}|$ 公式1-2；

[0093] 切比雪夫距离则适用于度量一些极端情况下的准标识符距离，它等于元组在各个维度上差异的最大值。用切比雪夫距离计算元组$i=(x_{i1},x_{i2},\cdots,x_{in})$和$j=(x_{j1},x_{j2},\cdots,x_{j2})$的距离公式可表示为：

[0094] $d(i,j)=\max(|x_{i1}-x_{j1}|,|x_{i2}-x_{j2}|,\ldots,|x_{in}-x_{jn}|)$ 公式1-3；

[0095] ②分类型准标识符距离度量

[0096] 分类型准标识符的距离度量需要借助相应的泛化层次结构树距离。如图1所示为本发明实施例提供的一个分类型准标识符“Marital Status”(婚姻状况)的泛化层次结构树。该准标识符包含7种不同的取值，最多能够泛化2次，因此准标识符“Marital Status”的泛化层次结构树包含2个泛化层次、7个叶子节点(Never-marrie,Married-civ-spous,Married-AF-spouse,Divorced,Separate,Widowed,Married-spouse-absent)，根节点为“*”。在泛化层次结构树中，每个父节点包含一个或多个子节点，没有子节点的节点称为叶子节点，而每个叶子节点表示一种分类型准标识符的取值。分类型准标识符的距离度量方法如下：

[0097] 当比较准标识符值“Divorced”和“Separate”的距离时，首先需要找到两个准标识符值的最小公共祖先，为“leave”。此时可以从纵向或横向的维度计算“Divorced”和“Separate”的距离。纵向维度是指通过比较子树与泛化层次结构树的总高度的比值计算距离。“Divorced”和“Separate”的最小公共祖先的子树高度为1，树的总高度为2，因此纵向维度的距离为1/2。横向维度则是通过比较子树与泛化层次结构树的叶子节点数量的比值计算距离。“Divorced”和“Separate”的最小公共祖先的子树的叶子节点数量为2，树的总叶子节点数量为7，因此横向维度的距离为2/7。由此可知，横向维度距离与叶子节点的数量相关性更强，而纵向维度距离则与泛化层次结构树高度相关性更强。因此当准标识符的取值数较多时，MSAK匿名算法使用横向维度距离度量分类型准标识符的距离；当分类型准标识符的泛化层次结构树的高度较高时，MSAK匿名算法使用纵向维度距离度量分类型准标识符的距离。

[0098] (3)敏感属性

[0099] 疾病诊断类属性是医学数据中最常见的敏感属性。因为疾病诊断属性属于多语义分类属性,因此在匿名化处理过程中进行ℓ-Diversity模型判断时,需要考虑疾病诊断的所有分类情况,才能够避免低估不同疾病诊断之间的相似性,降低相似性攻击的风险。MSAK匿名算法通过敏感属性语义标准化归并的过程,为疾病诊断属性构建多语义分类树,也为后续多语义敏感属性差异性计算奠定基础。

[0100] 具体的,敏感属性语义标准化归并过程:敏感属性的取值具有多种表现形式,如ICD编码、诊断名称、诊断编码和疾病二元取值等都属于疾病诊断属性。为了能够计算这些不同形式的疾病诊断的相似性,MSAK匿名算法对所有相关属性进行了语义标准化归并,将所有属性值都映射到MeSH(医学主题词表,Medical Subject Headings)或CMeSH(中文医学主题词表,Chinese Medical Subject Headings)的实体中。然后基于每个MeSH或CMeSH疾病实体的树编码,构建相应的多语义分类树,并利用多语义分类树来计算疾病诊断值属性的最小差异性。

[0101] 使用MeSH或CMeSH作为疾病诊断属性语义标准化归并的对象具有两方面的优点。一是MeSH和CMeSH中每个实体都有多个款目词,可以提高疾病诊断属性映射的准确率;二是MeSH和CMeSH本身就是被广泛应用的多语义分类系统,其分类的权威性保证了后续基于多语义分类树计算的最小差异性结果具有可靠性。

[0102] 其中,(1)面向抵御相似性攻击构建敏感属性多语义分类树,如图2所示为疾病诊断值为“新型冠状病毒感染”时,通过CMeSH构建的多语义分类树示例图。从图2中可以看出,“新型冠状病毒感染”有5种语义分类情况,因此在多语义分类树的5条主题词树路径中出现,树路径的深度不同,最大为6,最小为3。通过对每个敏感属性值的多语义分类树进行整合,就能够得到整体的敏感属性多语义分类树,进一步实现敏感属性最小差异性的计算。

[0103] (2)基于多语义分类树的敏感属性最小差异性计算:每个多语义敏感属性的分类树都有多个树路径,计算两个不同多语义敏感属性值的最小差异性相当于寻找两个属性值的距离最近的两条树路径。下面以基于多语义分类树计算慢性阻塞性肺疾病(chronic obstructive pulmoriary disease,COPD)和慢性肾衰竭(Chronic Kidney Failure,CKF)的最小差异性作为例,具体计算过程如下:

[0104] COPD在MeSH中有2条主题词树路径,因此树编码有2个,分别为C08.381.495.389、C23.550.291.500.875;CKF在MeSH中有4条主题词树路径,因此树编码有4个,分别为C12.200.777.419.780.750.500、C12.950.419.780.750.500、C23.550.291.500.906.500、C12.050.351.968.419.780.750.500。通过对两个实体的每个树编码进行两两自上而下的层序遍历比较可得,COPD的树编码C23.550.291.500.875与CKF的树编码C23.550.291.500.906.500的距离最近。这两个树编码最大深度为5,两个实体与最小公共父类构建的子树的深度为2,则此时可以参考分类型准标识符的纵向维度距离度量方法计算得到两者的距离为2/5,即COPD和CKF的最小差异性为2/5。

[0105] 具体的,通过第一次聚类可以降低数据规模,提高执行效率;数据集聚类划分过程:数据集聚类划分过程通过第一次聚类,将原始数据集划分成n个大小相近的第一子数据集,数据集聚类划分过程伪代码如表1-1所示。

[0106] 表1-1

输入：由 m 个准标识符 $\{R_1, R_2, \ldots, R_m\}$ 构成的数据集 S，数据集划分参数 n，敏感属性多语义分类树

输出：n 个第一子数据集 $\{S_1, S_2, \ldots, S_n\}$，敏感属性差异性矩阵

1. 对于每个准标识符 $R_i \in \{R_1, R_2, \ldots, R_m\}$：
2. 统计各取值出现的频率
3. 将属性值出现频率升序排列

[0107]

4. 等间距选取 n 个值，各准标识组合构建 n 个虚拟初始聚类中心
5. 对于每个元组 $s_i \in S$：
6. 将元组 s_i 与所有聚类中心进行距离比较
7. 将元组归到最近的聚类簇中，并更新该聚类簇的聚类中心
8. 记录非重复敏感属性值，计算两两敏感属性值的差异性，构建差异性矩阵
9. return $\{S_1, S_2, \ldots, S_n\}$，敏感属性差异性矩阵

[0108] 从表1-1中可以看到，该过程的输入包括结构化医学数据集、需要划分的第一子数据集数量n以及敏感属性语义标准化归过程构建的敏感属性多语义分类树，经过算法各步骤后，输出n个第一子数据集和敏感属性差异性矩阵。结构化医学数据集如表1所示。

[0109] 表1

[0110]

ID	性别	出生日期	邮编	医保类型	疾病诊断
1	女	1980.11.12	100010	城镇职工基本医疗保险	新型冠状病毒感染
2	女	1980.11.3	100010	城镇职工基本医疗保险	新型冠状病毒感染
3	男	1980.11.1	100020	新型农村合作医疗	肺脓肿
4	女	1980.11.30	100010	城镇居民基本医疗保险	肺脓肿
5	男	1980.11.11	100020	新型农村合作医疗	胸膜炎

[0111]

...

[0112] 数据集聚类划分过程的步骤1～4的目的是为了构建虚拟初始聚类中心，提高聚类生成的第一子数据集内元组的相似性。这是针对传统基于聚类的匿名化算法受初始聚类中心的影响较大，随机选取聚类中心会导致第一子数据集划分效果不稳定的问题进行的改进。构建虚拟初始聚类中心的流程如图3所示，具体流程描述如下：

[0113] 首先，统计结构化医学数据集每个准标识符各取值的频率，然后数值型准标识符(出生日期、邮编)依据数值从小到大升序排列，而分类型准标识符(性别、医保类型)依据各属性值出现频率从小到大按比例进行升序排列。然后根据需要划分的第一子数据集数量，对各序列进行相应等分(如图3所示为2等分时的示例)，然后每个部分的中线对应的各序列的值即为虚拟初始聚类中心。

[0114] 数据集聚类划分过程的步骤5～7表示，在构建虚拟初始聚类中心后，将所有元组逐个与这些聚类中心进行距离比较，将每个元组纳入到距离最近的聚类簇中。每次纳入元组后都更新聚类簇的聚类中心，同时距离受聚类簇的大小加权，控制第一子数据集的大小，

保证最终生成的n个第一子数据集大小相近。

[0115] 因为上述过程遍历了每一个元组，因此在步骤8中，可以记录得到结构化医学数据集中所有不重复的敏感属性值。然后结合敏感属性语义标准化归并过程构建的敏感属性多语义分类树，可以计算得到敏感属性差异性矩阵，用于后续过程的ℓ-Diversity模型的判断。如图4所示为敏感属性差异性矩阵示例图，矩阵中每个值为两种疾病对应的最小差异性。

[0116] 具体的，离群值检测分离过程：离群值检测分离过程在数据集聚类划分过程生成的所有第一子数据集中进行。分离离群值可以采用孤立森林算法、LOF算法或者其他可以进行离群值检测的算法。因为孤立森林算法时间复杂度较低，可以减少离群值检测分离过程对匿名化算法执行效率的影响，所以MSAK匿名算法通过孤立森林算法检测这些第一子数据集的离群值并将其分离，离群值检测分离过程伪代码如表1-2所示。

[0117] 表1-2

[0118]

输入：n个第一子数据集$\{S_1, S_2, \ldots, S_n\}$，离群值比例参数o 输出：n个去除离群值的第二子数据集$\{S'_1, S'_2, \ldots, S'_n\}$，第一待分配元组集合

[0119]

1. 对于每个第一子数据集$S_i \in \{S_1, S_2, \ldots, S_n\}$： 2. 根据泛化层次结构树将第一子数据集的分类属性值转化为哑变量值 3. 利用scikit-learn模块的孤立森林算法拟合数据，离群值比例设定为o 4. 生成离群值集合$\{O_1, O_2, \ldots, O_n\}$，纳入到第一待分配元组集合中 5. 从第一子数据集S_i中剔除离群值$\{O_1, O_2, \ldots, O_n\}$ 6. return $\{S'_1, S'_2, \ldots, S'_n\}$，第一待分配元组集合

[0120] 从表1-2中可以看到，该过程的输入包括所有第一子数据集、离群值比例参数o，经过算法各步骤后，输出分离了离群值的第二子数据集以及由离群值构成的第一待分配元组集合。

[0121] 离群值检测分离过程的步骤1表示该算法过程在各第一子数据集中独立进行，互不干扰。因此在本过程中可以通过对各第一子数据集进行并行处理，提高算法执行效率。步骤2～5是使用孤立森林算法检测分离离群值的过程，具体流程如图5所示。作为一种常用于检测离群值的算法，孤立森林算法在许多编程语言中都有成熟的库或框架可供使用，如Python的scikit-learn、PyOD;Java的Weka、R的isolationForest以及MATLAB的Isolation Forest Toolbox等。孤立森林算法的具体原理描述如下：

[0122] 首先随机选择数据集的一个特征，并在该特征的最大值和最小值之间随机选择1个分割阈值。然后将各元组根据其在该特征上的取值分成两个子集(左子集和右子集)，将每个子集作为1个新的节点，并在子集中重复上述过程，直到每个子集只剩余1个样本点，即只包含1个元组为止。这个过程可以形成一个二叉树结构，其中根节点表示整个数据集，叶节点表示单个元组。孤立森林算法通过重复执行上述步骤构建多棵随机二叉树。对于每个元组，算法可以计算出它被隔离的平均深度。由于异常值在数据集中更容易被分隔开来，它们通常比正常值需要更少的深度才能被隔离。因此，具有较小平均隔离深度的数据点可以

被认为是异常值。

[0123] 需要注意的是，离群值比例参数的确定是影响离群值检测分离结果的重要因素。一些文献提到，数据集的离群值比例通常在1％到10％之间，但也有些情况，尤其是一些高维度的数据集，离群值比例会高于10％。在MSAK匿名算法整个过程中，检测分离离群值的目的是减少离群值对聚类结果的影响，离群值的存在会导致整个聚类簇都产生过度的泛化，因此离群值比例参数设定在较高取值时，整体聚类效果会更好。同时，离群值比例参数的确定还需要结合医学数据准标识符的特点来确定，当医学数据集中存在较多的异质性较大的准标识符时，可以适当提高离群值比例参数；当医学数据集中存在较多的时间类等异质性较小的准标识符时，可以适当减小离群值比例参数。

[0124] 具体地，子数据集再聚类过程：子数据集再聚类过程是MSAK匿名算法的第二次聚类过程，该过程不仅将去除了离群值后的第二子数据集的元组聚集成簇，而且还对这些聚类簇进行了裁剪和判断，并标记满足ℓ-Diversity模型的情况。子数据集再聚类过程伪代码如下表1-3所示：

[0125] 表1-3

[0126]

```
输入：分离离群值后的第二子数据集{S'1,S'2,...,S'n}，第一待分配元组集合,隐私保护模型参数k和参数ℓ
输出：第一聚类簇集合S"1，第二聚类簇集合S"2，第二待分配元组集合
1. 对于每个第二子数据集S'i∈{S'1,S'2,...,S'n}：
2.      在第二子数据集S'i中随机选取f个元组作为聚类中心
3.      对于剩余的每个元组si∈ S'i：
4.          将元组si与所有聚类中心进行距离比较
5.          将元组归到最近的聚类簇中并更新该聚类簇的聚类中心
6.      对于每个聚类簇 Cj ∈ S'i
7.          if 聚类簇的大小<k：
8.              将聚类簇Cj纳入到过第一聚类簇集合S"1中
9         else:
10.             if 聚类簇的大小>k：
11.                 剔除距离过远元组纳入到第一待分配元组集合中，仅保留k个元组
12.             if 聚类簇满足ℓ-Diversity模型：
13.                 将聚类簇Cj纳入到第二聚类簇集合S"2中
14.           else:
15                  将聚类簇Cj纳入到第一聚类簇集合S"1中
16.return  第一聚类簇集合S"1，第二聚类簇集合S"2，第二待分配元组集合
```

[0127] 从表1-3中可以看到，该过程的输入包括所有分离了离群值的第二子数据集、第一待分配元组集合、k-Anonymity模型的参数k和ℓ-Diversity模型的参数ℓ，经过算法各步骤后，输出第一聚类簇集合$S"_1$，第二聚类簇集合$S"_2$和第二待分配元组集合。其中，$f=\frac{g}{k}$；g表示

子数据集的元组数。

[0128] 与离群值检测分离过程相似，子数据集再聚类过程的步骤1也表示该算法过程在各第二子数据集中独立进行，互不干扰，可以通过并行计算的方法提高算法执行效率。步骤2表示在第二子数据集中随机选取了$\frac{\text{子数据集的元组数}}{k}$(向下取整)个元组作为聚类中心，然后在步骤3～5中，将剩余元组逐个分配到同一个第二子数据集内距离最近的聚类簇中，每次分配都更新聚类中心。当所有元组都分配完毕后，每个第二子数据集便生成了$\frac{\text{子数据集的元组数}}{k}$(向下取整)个大小不一的聚类簇。

[0129] 步骤6～15为后续对这些聚类簇的裁剪和ℓ-Diversity模型判断过程，具体流程如图6所示。在该过程中，首先判断每个聚类簇的大小，如果聚类簇大小小于k，则将聚类簇纳入到第一聚类簇集合中；如果聚类簇的大小大于k，则保留k个元组，将距离较远的元组分离出来，判断剩余元组是否满足ℓ-Diversity模型，如满足则将剔除过远元组后的聚类簇纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中；将分离的元组纳入到第二待分配元组集合中；如果聚类簇的大小等于k，则直接判断聚类簇是否满足ℓ-Diversity模型，如满足则纳入到第二聚类簇集合中，否则纳入到第一聚类簇集合中。

[0130] 由于MSAK匿名算法在判断聚类簇是否满足ℓ-Diversity模型时考虑到了医学数据集中敏感属性的多语义性特点，因此不能像传统的基于聚类的匿名化算法一样只是简单地计算聚类簇中出现互不相同的敏感属性值数量是否到阈值ℓ，而是需要判断各敏感属性值相互之间的差异性之和是否达到阈值ℓ。敏感属性差异性判断流程伪代码如表1-4所示：

[0131] 表1-4

输入：聚类簇C_j的h个敏感属性值$\{l_1, l_2, \ldots, l_h\}$，敏感属性差异性矩阵，隐私保护模型参数$\ell$
输出：布尔值

[0132]

1. 读取数据集划分过程构建的敏感属性差异性矩阵
2. 计算聚类簇C_j中的敏感属性值的整体差异性为$D=\frac{\sum_{i=1}^{h}\sum_{j=1,j\neq i}^{h} l_i l_j}{h(h-1)}$
3. If $D \geq \frac{\ell-1}{h-1}$:
4. return True
5. else:
6. return False

[0133] 该判断流程输入聚类簇的各敏感属性值、敏感属性差异性矩阵和ℓ-Diversity模型的参数ℓ，经过算法各步骤后，输出判断结果布尔值。步骤1读取数据集聚类划分过程构建的敏感属性差异性矩阵，减少了对敏感属性差异性的重复计算。步骤2计算整个聚类簇所有敏感属性的整体差异性，计算过程如下：

[0134] 假设聚类簇中共有h个敏感属性值，分别为$l_1, l_2, \ldots, l_h$，则聚类簇的整体敏感属性差异性等于各敏感属性值相互间的差异性之和，用公式表示为：

[0135] $$d = \sum_{i=1}^{h} \sum_{j=1,j\neq i}^{h} l_i l_j \qquad \text{公式 1-4;}$$

[0136] 为了使不同大小的聚类簇能够用同一个标准进行ℓ-Diversity判断，还需要对该结果进行归一化处理。由于聚类簇中有h个元素，而计算整体差异性时每个元素都需要比较h-1次，因此敏感属性差异性比较次数为h*(h-1)，因此聚类簇整体差异性归一化公式为：

[0137] $$D = \frac{d}{h(h-1)} \qquad \text{公式 1-5;}$$

[0138] 该判断流程的步骤3～6为判断聚类簇的整体敏感属性差异性是否满足ℓ-Diversity模型阈值的过程。由于在计算整体差异性时，聚类簇中每个元素都需要比较h-1次，因此只需要保证在这h-1次中，有ℓ-1次的差异性结果为1，便能满足ℓ-Diversity模型。用公式表达为：

[0139] $$D_{min} = \frac{h(l-1)}{h(h-1)} = \frac{\ell-1}{h-1} \qquad \text{公式 1-6;}$$

[0140] 聚类簇的整体敏感属性差异性越大，抵御相似性攻击的能力也越强。因此当聚类簇的整体差异性D达到D_{min}及以上时，可以认为聚类簇中敏感属性值的差异性已满足ℓ-Diversity模型。

[0141] 具体的，聚类簇合并调整过程：聚类簇合并调整过程通过将第二待分配元组逐个分配到各聚类簇中，然后通过聚类中心取代聚类簇内所有元组的方式实现匿名化的过程。聚类簇合并调整过程伪代码如下表1-5所示：

[0142] 表1-5

[0143]

输入：第一聚类簇集合S”$_1$，第二聚类簇集合S”$_2$，第二待分配元组集合，隐私保护模型参数k和参数ℓ
输出：匿名化医学数据集
1. 对于每个元组s_i∈第二待分配元组集合：
2. if 第一聚类簇集合S”$_1$大小>0：
3. 将元组s_i与第一聚类簇集合S”$_1$中各聚类中心比较距离，距离最

近的簇为 C_{close}

4. if 元组 s_i 纳入聚类簇 C_{close} 后满足ℓ-Diversity 模型:

5. 将元组 s_i 纳入聚类簇 C_{close}，并更新聚类中心

6. if 纳入元组后的 C_{close} 大小≥k:

7. 将聚类簇 C_{close} 从第一聚类簇集合 S''_1 中排除并纳入到第二聚类簇集合 S''_2

8. else:

[0144] 9. 将元组 s_i 与第二聚类簇集合 S''_2 中各簇的聚类中心比较距离，并纳入到最近的簇中

10. else:

11. 将元组 s_i 与第二聚类簇集合 S''_2 中各簇的聚类中心比较距离，并纳入到最近的簇中

12. 合并调整第一聚类簇集合 S''_1，用聚类中心取代第二聚类簇集合 S''_2 内所有元组，形成匿名化医学数据集

13. return 匿名化医学数据集

[0145] 从表1-5中可以看到，该过程的输入包括第一聚类簇集合S''_1，第二聚类簇集合S''_2和第二待分配元组集合、k-Anonymity模型的参数k和ℓ-Diversity模型的参数ℓ，经过算法各步骤后，输出匿名化医学数据集，如表2所示。

[0146] 表2

[0147]

ID	性别	出生日期	邮编	医保类型	疾病诊断
*	*	1980.11	100010，100020	社会医疗保险	新冠
*	*	1980.11	100010，100020	社会医疗保险	新冠
*	*	1980.11	100010，100020	社会医疗保险	肺脓肿
*	*	1980.11	100010，100020	社会医疗保险	肺脓肿
*	*	1980.11	100010，100020	社会医疗保险	胸膜炎
…					

[0148] 聚类簇合并调整过程的步骤1～11是将第二待分配元组逐个分配到第一聚类簇集合S''_1，第二聚类簇集合S''_2中所有聚类簇的分配过程，也是MSAK匿名算法的第三次聚类过程，具体流程如图7所示。该分配过程可以定义3个分配原则，分别为基于最小簇长约束原则、基于满足差异性约束原则和基于最小信息损失原则。基于最小簇长约束原则是指当第一聚类簇集合S''_1的大小大于0时，优先将待分配元组分配到第一聚类簇集合中，确保每个聚类簇最小簇长都能达到k，以满足k-Anonymity模型的要求；基于满足差异性约束原则是指所述第一聚类簇集合中的聚类簇纳入所述待分配元组集合后满足ℓ-Diversity模型，才能从第一聚类簇集合中剔除并纳入到所述第二聚类簇集合中，确保所有聚类簇最终都能满足ℓ-Diversity模型；基于最小信息损失原则是指待分配元组只分配到距离最近的聚类簇中，对剩余的第一聚类簇集合中的聚类簇进行就近合并，直至同时满足k-Anonymity模型和ℓ-Diversity模型后纳入第二聚类簇集合，最后仍不满足者则抑制处理，尽可能地减少信息损失。

[0149] 聚类簇合并调整过程的步骤12合并调整第一聚类簇集合S''_1，用聚类中心取代第

二聚类簇集合S”$_2$内所有元组，使每个聚类簇分别生成一个等价类，每个等价类由多条相同元组构成，大小等于相应聚类簇的大小，所有等价类共同构成了匿名化医学数据集。

[0150] 在本发明的一个具体实施例中，如图10所示，匿名化医学数据集满足k-Anonymity模型和ℓ-Diversity模型，可以抵御各种类型的攻击，包括链接攻击、同质性攻击、偏斜攻击和相似性攻击。从执行效率、信息损失和隐私泄露风险三个维度进行匿名化效果评估；所述执行效率可以通过算法运行时间进行评估；所述信息损失可以通过抑制率、标准化确定性惩罚、信息损失比率和/或非均衡熵等方面进行评估；所述隐私泄露风险可以通过链接攻击风险、同质性攻击风险、偏斜攻击风险和/或相似性攻击风险等方面进行评估。用于平衡数据可用性与隐私性；进而将数据作为共享医学数据。若要求较高隐私保护水平，可以设置k>=5，ℓ>=2。

[0151] 本发明采用MSAK匿名算法对医学数据集进行匿名化处理的效果可以从算法执行效率、信息损失以及隐私披露风险3个维度进行。匿名化结果平衡了数据安全性和可用性，能够满足数据共享者和隐私相关政策的要求及研究使用者的需要，实现了医学数据的隐私共享保护。如图8所示为本发明实施例提供的处理前包含隐私信息的数据示意图。如图9所示为本发明实施例提供的处理后的数据隐私信息被泛化并可用于数据分析的示意图。

[0152] 在本发明的另一个具体实施例中提供了三种代表性的基于聚类的匿名化算法(KNN算法、k-member算法、OKA算法)和一种目前性能领先的全局泛化算法(FLASH算法)。

[0153] (1)KNN算法

[0154] Knn算法的核心是随机选取聚类中心，然后依次选取最邻近的k-1个元组聚集成簇。因此，只需要保证每次生成的k个元组的敏感属性值不完全相同，即可实现l＝2的ℓ-Diversity模型。具体实现方法是，首先选取最邻近的k-2个元组聚集成簇，判断纳入第k-1个元素时是否能满足ℓ-Diversity模型，如果满足则纳入，不满足则判断下一个最邻近元素，直至能满足条件后纳入满足条件的元组形成聚类簇。

[0155] (2)k-member算法

[0156] k-member算法与Knn算法最大的区别是，k-member算法选取邻近元组是逐个进行的，需要不断更新聚类中心。因此可以采用相似的思路，通过判断纳入第k-1个元组后是否能满足ℓ-Diversity模型，如果满足则纳入，不满足则判断下一个最邻近元素，直至能满足条件后纳入满足条件的元组形成聚类簇。

[0157] (3)OKA

[0158] OKA由聚类阶段和调整阶段两个阶段构成，在调整阶段中，需要逐个将多余元组与各聚类簇进行距离比较。因此只需要在聚类阶段将未满足ℓ-Diversity模型的聚类簇标记出来，然后在调整阶段逐个分配多余元组时，优先把能够让未满足ℓ-Diversity模型的聚类簇实现ℓ-Diversity模型的元组分配给这些聚类簇，便能实现l＝2的ℓ-Diversity模型保护。

[0159] 本发明实施例选取了三种代表性的基于聚类的匿名化算法(kNN算法、k-member算法、OKA算法)和一种全局泛化算法(FLASH算法)作为比较算法，通过使用adult数据集和MIMIC-IV的疾病诊断字段构建仿真实验数据集进行了仿真实验验证。在相同实验条件下，比较了MSAK匿名算法和其他算法匿名化结果的执行效率、信息损失及隐私披露风险。结果表明在较高隐私保护水平条件下，较大规模的医学数据匿名化处理时，MSAK匿名算法的执

行效率高于其他基于聚类的匿名化算法;抑制率和整体信息损失表现都优于所有其他算法;而且还能够明显降低链接攻击风险和相似性攻击风险,能够较好平衡数据安全性和可用性。

[0160] 本发明识别和归并需要进行隐私保护的属性,明确匿名化处理对象;采用MSAK匿名算法第一次聚类将原始数据集划分为多个第一子数据集,分离每个第一子数据集的离群值,形成第二子数据集和第一待分配元组集合;第二次聚类对每个第二子数据集聚类,生成多个由相似元组构成的聚类簇,判断聚类簇的大小和过远元组剪裁,形成第一聚类簇集合、第二聚类簇集合和第二待分配元组集合;在所有聚类簇的基础上进行第三次聚类,将所有待分配元组分配到第一聚类簇集合、第二聚类簇集合中,用聚类中心取代聚类簇内所有元组,生成满足隐私保护模型的等价类,从而实现数据的匿名化

[0161] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0162] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

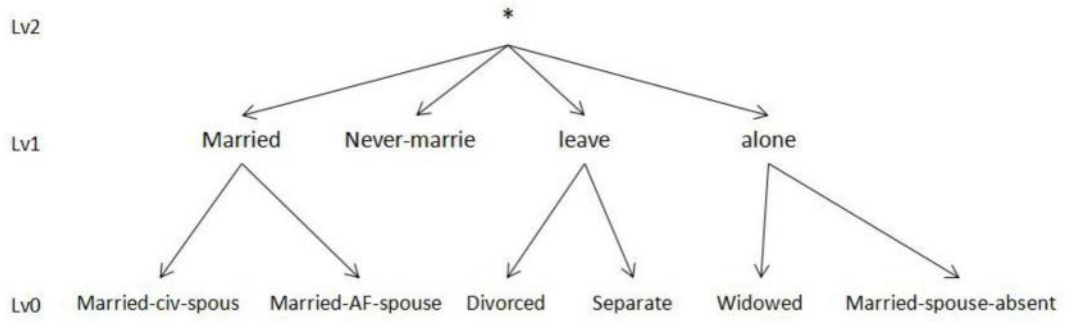

Lv2
*
Lv1
Married
Never-marrie
leave
alone
Lv0
Married-civ-spous
Married-AF-spouse
Divorced
Separate
Widowed
Married-spouse-absent

图1

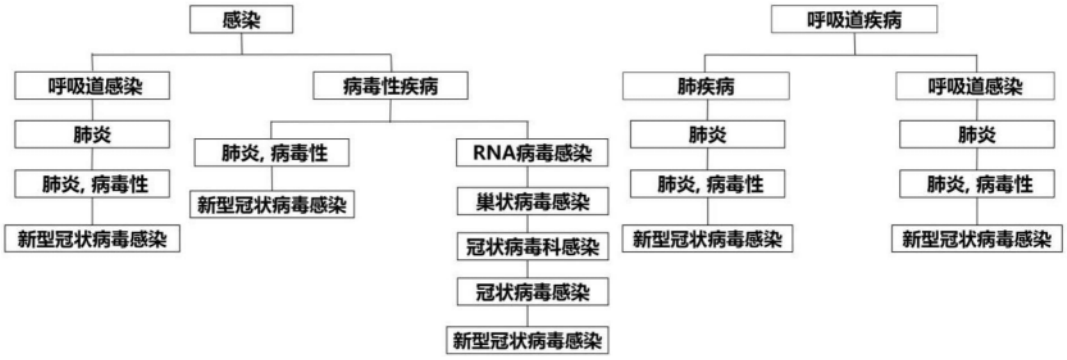

感染
呼吸道感染
肺炎
肺炎，病毒性
新型冠状病毒感染
病毒性疾病
肺炎，病毒性
新型冠状病毒感染
RNA病毒感染
巢状病毒感染
冠状病毒科感染
冠状病毒感染
新型冠状病毒感染
呼吸道疾病
肺疾病
肺炎
肺炎，病毒性
新型冠状病毒感染
呼吸道感染
肺炎
肺炎，病毒性
新型冠状病毒感染

图2

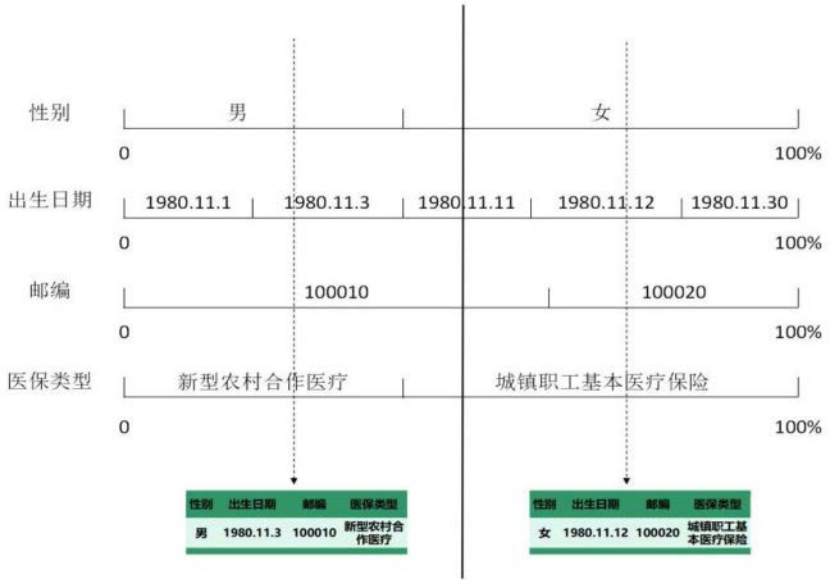

性别
男
女
0
100%
出生日期
1980.11.1
1980.11.3
1980.11.11
1980.11.12
1980.11.30
0
100%
邮编
100010
100020
0
100%
医保类型
新型农村合作医疗
城镇职工基本医疗保险
0
100%
性别 出生日期 邮编 医保类型
男 1980.11.3 100010 新型农村合作医疗
性别 出生日期 邮编 医保类型
女 1980.11.12 100020 城镇职工基本医疗保险

图3

	新型冠状病毒感染	肺脓肿	胸膜炎	感冒	严重急性呼吸综合征
新型冠状病毒感染	0	0.75	0.75	0.75	0.17
肺脓肿	0.75	0	0.5	0.5	0.5
胸膜炎	0.75	0.5	0	0.5	0.5
感冒	0.75	0.5	0.5	0	0.5
严重急性呼吸综合征	0.17	0.5	0.5	0.5	0

图4

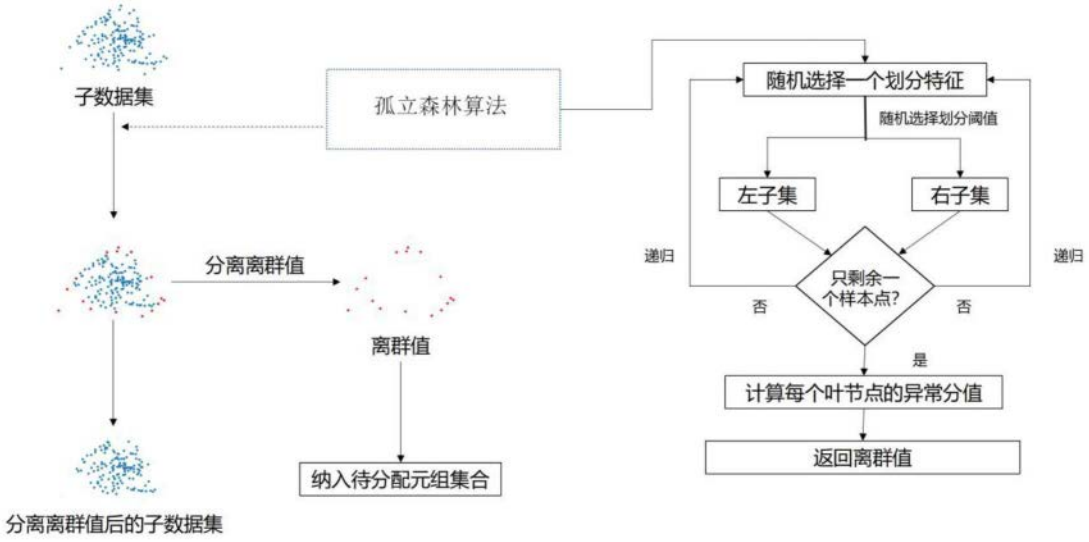


图5

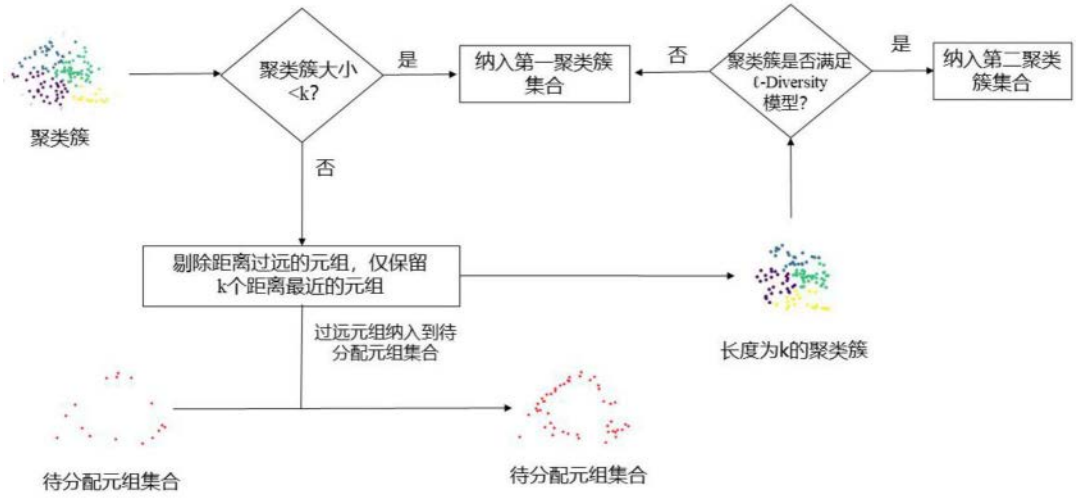

聚类簇
聚类簇大小<k?
是
纳入第一聚类簇集合
否
聚类簇是否满足ℓ-Diversity模型?
是
纳入第二聚类簇集合
否
剔除距离过远的元组，仅保留k个距离最近的元组
过远元组纳入到待分配元组集合
长度为k的聚类簇
待分配元组集合
待分配元组集合

图6

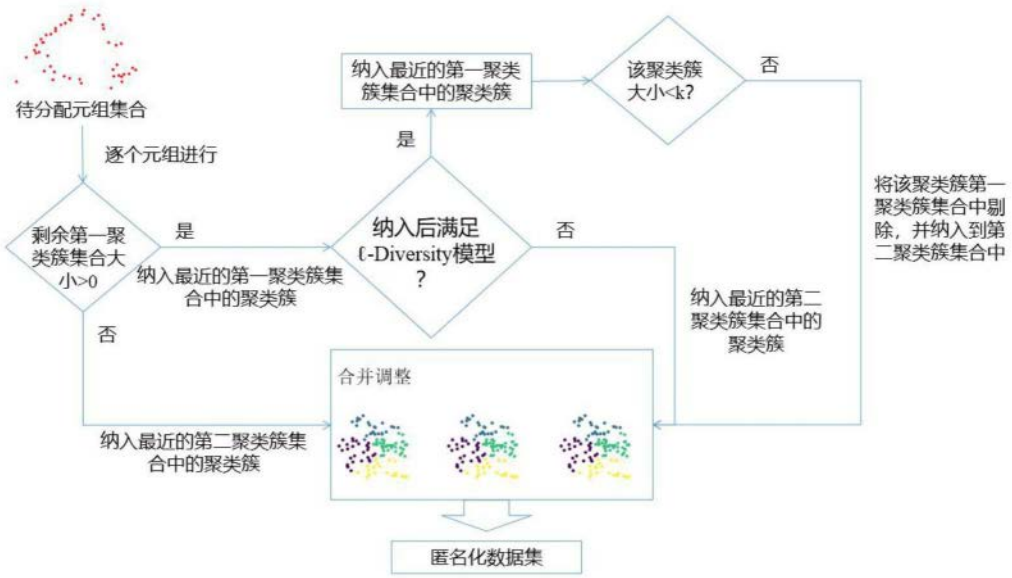

待分配元组集合
逐个元组进行
剩余第一聚类簇集合大小>0
是
纳入最近的第一聚类簇集合中的聚类簇
否
纳入最近的第二聚类簇集合中的聚类簇
纳入后满足ℓ-Diversity模型?
是
纳入最近的第一聚类簇集合中的聚类簇
该聚类簇大小<k?
否
将该聚类簇第一聚类簇集合中剔除，并纳入到第二聚类簇集合中
合并调整
匿名化数据集

图7

	age	workclass	education-num	marital-status	occupation	race	sex	native-country	disease (ICD10)
2	40	Private	13	Never-married	Prof-specialty	White	Female	United-States	G4700
3	23	Private	9	Never-married	Adm-clerical	White	Female	United-States	E785
4	52	Private	12	Married-civ-spouse	Exec-managerial	White	Male	United-States	D649
5	35	Private	14	Married-civ-spouse	Prof-specialty	White	Female	United-States	E785
6	21	Private	10	Never-married	Sales	White	Female	United-States	N183
7	18	Private	10	Never-married	Other-service	White	Female	United-States	D62
8	48	Private	16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	Male	India	F419
9	22	Private	9	Divorced	Machine-op-inspct	White	Female	United-States	N390
10	26	Private	10	Married-civ-spouse	Adm-clerical	White	Male	United-States	E875
11	39	Private	10	Married-civ-spouse	Sales	White	Male	Philippines	I480
12	27	Private	6	Married-civ-spouse	Craft-repair	White	Male	United-States	K219
13	53	Local-gov	10	Married-civ-spouse	Protective-serv	White	Male	United-States	E871
14	19	Private	10	Never-married	Handlers-cleaners	White	Male	United-States	K219
15	33	Private	15	Married-civ-spouse	Prof-specialty	White	Male	United-States	I10
16	39	Private	8	Divorced	Sales	White	Female	United-States	F419
17	36	Private	7	Never-married	Machine-op-inspct	White	Female	United-States	K219
18	40	Private	9	Married-civ-spouse	Sales	White	Male	United-States	E860
19	90	Private	4	Married-civ-spouse	Machine-op-inspct	White	Male	United-States	G4700
20	47	lf-emp-not-i	13	Divorced	Exec-managerial	Asian-Pac-Islander	Male	Japan	J449

图8

	age	workclass	education-num	marital-status	occupation	race	sex	native-country	disease (ICD10)
2	30, 41	*	14, 16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	*	China	G92
3	30, 41	*	14, 16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	*	China	I4891
4	30, 41	*	14, 16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	*	China	F419
5	30, 41	*	14, 16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	*	China	K219
6	30, 41	*	14, 16	Married-civ-spouse	Prof-specialty	Asian-Pac-Islander	*	China	F329
7	23, 76	Private	3, 5	Never-married	Handlers-cleane	White	Male	United-States	D696
8	23, 76	Private	3, 5	Never-married	Handlers-cleane	White	Male	United-States	G8929
9	23, 76	Private	3, 5	Never-married	Handlers-cleane	White	Male	United-States	N179
10	23, 76	Private	3, 5	Never-married	Handlers-cleane	White	Male	United-States	D649
11	23, 76	Private	3, 5	Never-married	Handlers-cleane	White	Male	United-States	I480
12	24, 44	gov	10, 13	Married-civ-spouse	*	Black	Female	*	I252
13	24, 44	gov	10, 13	Married-civ-spouse	*	Black	Female	*	K219
14	24, 44	gov	10, 13	Married-civ-spouse	*	Black	Female	*	N189
15	24, 44	gov	10, 13	Married-civ-spouse	*	Black	Female	*	N390
16	24, 44	gov	10, 13	Married-civ-spouse	*	Black	Female	*	F329
17	36, 63	Private	1, 4	*	*	Other	Male	*	I2510
18	36, 63	Private	1, 4	*	*	Other	Male	*	F17210
19	36, 63	Private	1, 4	*	*	Other	Male	*	I10
20	36, 63	Private	1, 4	*	*	Other	Male	*	I4891
21	36, 63	Private	1, 4	*	*	Other	Male	*	I10

图9

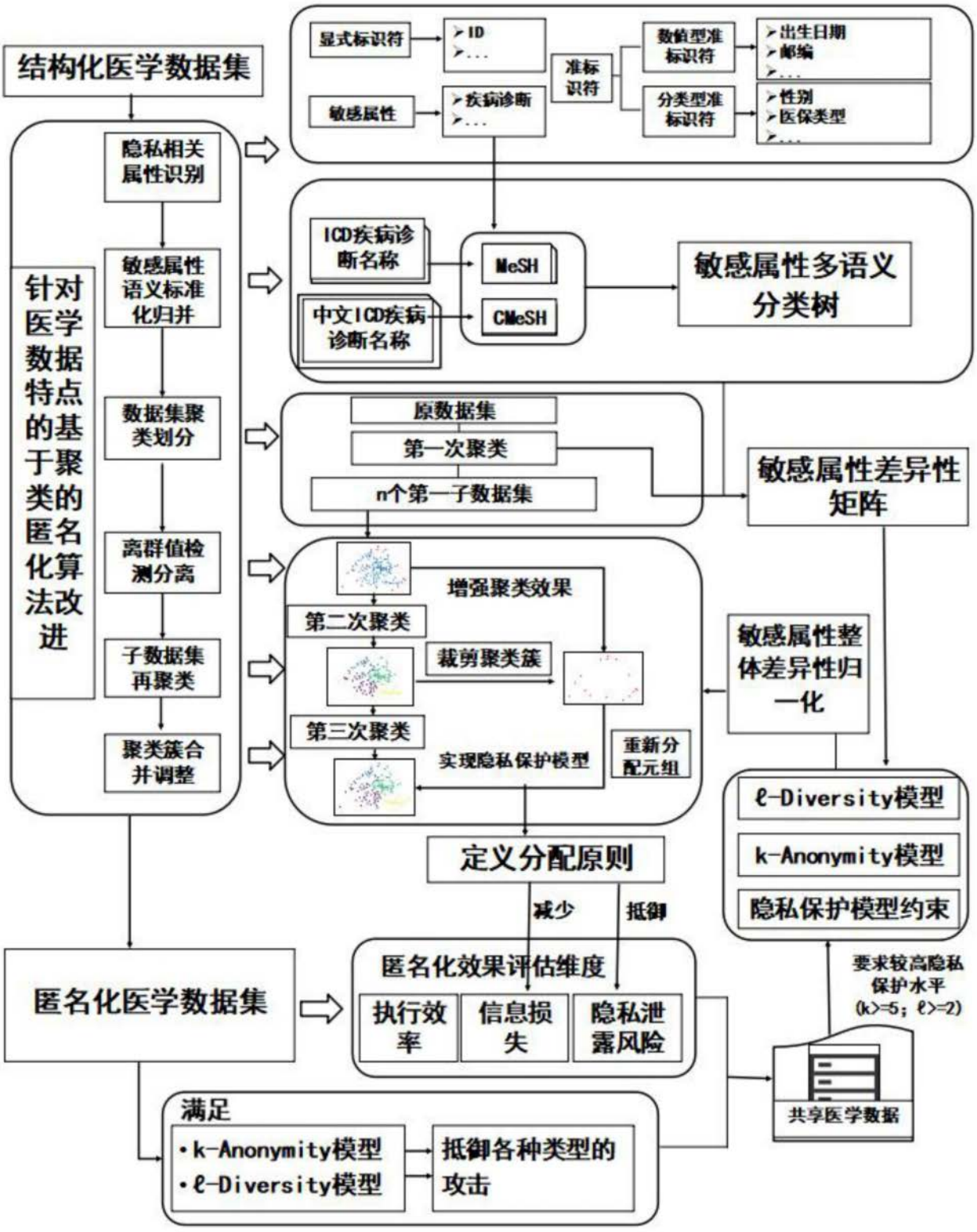

结构化医学数据集
显式标识符
ID
...
敏感属性
疾病诊断
...
准标识符
数值型准标识符
出生日期
邮编
...
分类型准标识符
性别
医保类型
...
针对医学数据特点的基于聚类的匿名化算法改进
隐私相关属性识别
敏感属性语义标准化归并
数据集聚类划分
离群值检测分离
子数据集再聚类
聚类簇合并调整
ICD疾病诊断名称
中文ICD疾病诊断名称
MeSH
CMeSH
敏感属性多语义分类树
原数据集
第一次聚类
n个第一子数据集
敏感属性差异性矩阵
增强聚类效果
第二次聚类
裁剪聚类簇
第三次聚类
实现隐私保护模型
重新分配元组
敏感属性整体差异性归一化
ℓ-Diversity模型
k-Anonymity模型
隐私保护模型约束
定义分配原则
减少
抵御
匿名化效果评估维度
执行效率
信息损失
隐私泄露风险
匿名化医学数据集
要求较高隐私保护水平
(k>=5; ℓ>=2)
共享医学数据
满足
•k-Anonymity模型
•ℓ-Diversity模型
抵御各种类型的攻击

图10