• 医学信息技术 •

医学数据共享隐私保护中基于聚类的匿名化算法关键技术研究*

唐明坤 吴思竹 周佳茵 段一凡 胡拯涌 钱 庆

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔摘要〕 **目的/意义** 基于聚类的匿名化算法具有灵活性较高、适用范围较广、能够保留原始数据更多信息的特点。合理使用基于聚类的匿名化算法进行匿名化处理可以获得满足隐私保护需求的高质量医学数据。**方法/过程** 通过文献调研法和比较分析法，梳理面向医学数据共享、基于聚类的匿名化算法关键技术，概述该类算法的主要流程，归纳与之相关的隐私模型，包括具有代表性的传统隐私模型和个性化隐私模型，并分析代表性算法的优点和不足。**结果/结论** 应当合理选择基于聚类的匿名化算法类型、灵活改进算法模型，加大算法工具研发力度，以推动医学数据安全便利和高质量共享。

〔关键词〕 数据共享；隐私保护；聚类算法；数据匿名化；隐私模型

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.06.011

Study on Key Technologies of Clustering－based Anonymization Algorithm in Medical Data Sharing Privacy Protection

TANG Mingkun, WU Sizhu, ZHOU Jiayin, DUAN Yifan, HU Zhengyong, QIAN Qing

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕 **Purpose/Significance** The anonymization algorithm based on clustering has the characteristics of high flexibility, wide application range, and the ability to retain more information of the original data. Reasonable use of clustering－based anonymization algorithm for anonymization can obtain high－quality medical data that meets the needs of privacy protection. **Method/Process** Through literature research and comparative analysis, the study sorts out the key technologies of clustering－based anonymization algorithm for medical data sharing, summarizes the main process of such algorithm and the related privacy models, including representative traditional privacy models and personalized privacy models, and analyzes the advantages and disadvantages of representative clustering－based anonymization algorithm. **Result/Conclusion** The type of clustering－based anonymization algorithms should be reasonably selected, the model of clustering－based anonymization algorithms should be flexibly improved, and the research and development of clustering－based anonymization tools should be heavily invested. These measures can promote safe, convenient and higher quality sharing of medical data.

〔**Keywords**〕 data sharing; privacy protection; clustering algorithm; data anonymization; privacy model

1 引言

随着数据共享需求的不断增长，数据隐私保护问题受到越来越多关注。近年来我国相继出台《中

〔修回日期〕 2022－12－20

〔作者简介〕 唐明坤，硕士研究生；通信作者：钱庆，研究员。

〔基金项目〕 国家重点研发计划（项目编号：2021YFC2701301）；中国医学科学院医学与健康科技创新工程项目（项目编号：2021－I2M－1－057）。

华人民共和国个人信息保护法》等法律法规，对数据隐私保护提出更高要求[1]。医学数据中包含大量个人隐私信息，随着跨单位合作的增加，医学数据共享需求也在不断增长[2]，相关数据共享平台[3-4]和机制[5-7]等研究受到广泛关注。然而，由于包含大量个人隐私信息和群体健康生理信息，医学数据的高质量开放共享面临诸多挑战[8]。

基于聚类的匿名化算法具有灵活性较高、能够保留原始数据更多信息的特点，被广泛应用于各种对数据质量要求较高的场景，尤其是医学数据共享[9-11]。但医学数据具有语义信息层次丰富、隐私保护需求多样化等特点，加上不同类型基于聚类的匿名化算法在时间复杂度、适用数据特点、适用场景等方面存在较大差异，导致在真实世界数据共享中如何选择合适的基于聚类的匿名化算法成为迫切需要解决的问题。因此本研究通过梳理面向医学数据共享的匿名化聚类算法关键技术相关内容，以期为相关研究提供参考。

2 基于聚类的匿名化算法主要流程

2.1 主要流程

医学数据类型丰富，包括电子病历数据、公共卫生数据等。基于聚类的匿名化算法首先根据数据特点选择合适的隐私模型，而后将原始数据集映射到特定度量空间中，再对空间中的数据进行聚类和泛化、抑制或扰动等匿名化处理以实现数据匿名化。本文归纳其主要流程，包括待处理数据属性类别确定、隐私模型选择、匿名化聚类算法选择与实现，见图 1。

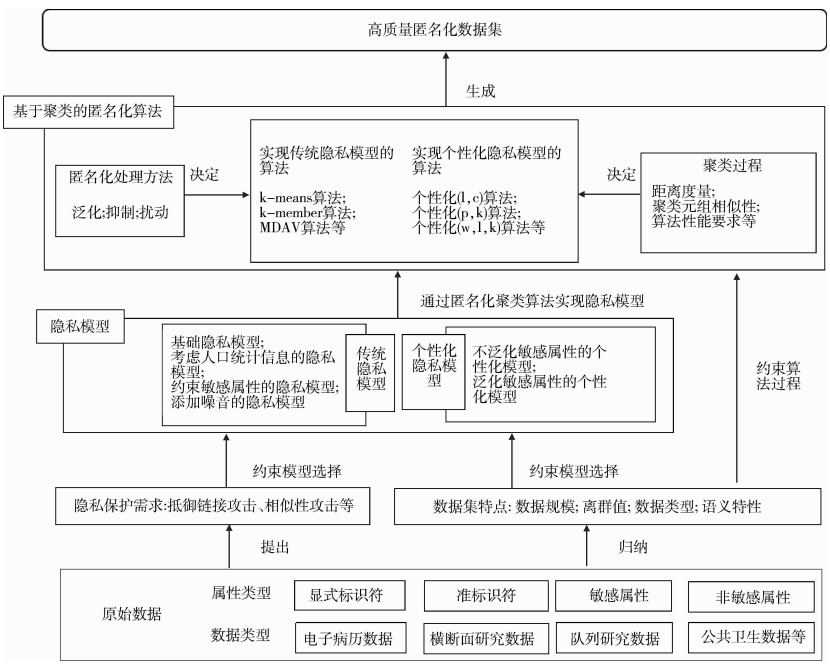

图 1 基于聚类的匿名化算法主要流程

2.2 数据属性类别确定

数据属性类型包括显式标识符、准标识符、敏感属性和非敏感属性 4 类。显式标识符指能直接确定个体身份的属性，如姓名等；准标识符指在一定背景知识下，能够通过该属性或属性组合确定个体身份的属性，如年龄等；敏感属性指需要保护、涉及个体隐私信息的属性，如疾病等；非敏感属性是不属于以上 3 类的属性。显式标识符直接暴露个体身份，需要进行抑制处理；而准标识符和敏感属性潜在暴露个体身份，是匿名化处理的重点对象。

2.3 隐私模型选择

隐私模型是数据集匿名化处理的标准。其可分

为传统隐私模型和个性化隐私模型。后者在传统隐私模型基础上对敏感属性值进行个性化保护，例如给予艾滋病患者相关数据更多保护。确定是否需要个性化保护后，还需要进一步综合隐私保护需求（如严格或宽泛程度）和数据特点（如数据规模、离群值数量等）。通常隐私保护需求越高则需要选择更严格的隐私模型。同时，数据集特点也约束着隐私模型的选择，例如大规模电子病历数据集，当数据量在总人口中占比较大时，需要使用考虑人口统计信息的隐私模型。

2.4 算法选择与实现

基于聚类的匿名化算法是实现隐私模型的方法。基于聚类的匿名化算法选择除了要考虑隐私模型的要求，还需要考虑算法的实现成本、数据集特点等。选取最佳匿名化算法才能够获得高质量匿名化数据集。不同算法之间的差别主要是聚类过程和匿名化处理。聚类过程指将数据集中相似元组聚集成簇的过程，该过程受距离度量方法影响，决定算法性能。匿名化处理指对每个聚类簇进行泛化、抑制或扰动从而使每个簇内的单个元组无法再与其他元组区别的过程。

3 隐私模型

3.1 传统隐私模型

3.1.1 k－匿名模型 经典传统隐私模型包括 k－匿名模型、l－多样性模型、t－近似性模型等，各模型要求及能抵御的攻击，见表 1。k－匿名模型是最早出现的隐私模型[12]，也是实现其他隐私模型的基础。该模型因为能够简单且有效降低重识别风险，至今仍被广泛应用于医疗数据、中医药临床数据等的隐私保护[13]。k－匿名模型能够抵御链接攻击，但存在同质性攻击风险，例如在一个医疗数据集中，当一个等价类中多个患者患有相同疾病，攻击者可以轻易确定满足该类准标识符的个体患有该疾病。

表 1 经典传统隐私模型的要求及可抵御的攻击风险

传统隐私模型	模型要求	抵御攻击
k－匿名模型[12]	要求匿名化数据中的每个等价类包含的相同元组数量不少于 k 个，将所有元组被重识别的概率降低到 1/k	链接攻击
l－多样性模型[14]	Distinct l－多样性：每个等价类中敏感属性值约束至少包含 l 个互不相同的取值	链接攻击、同质性攻击
	Entropy l－多样性：每个等价类中敏感属性值中信息熵约束至少为 log l	
	Recursive（c，l）－多样性：每个等价类中出现频率最高的敏感属性值的出现频率不高于第 l 项及之后所有的敏感属性值的出现频率之和与系数 c 的乘积	
	Recursive（c1，c2，l）－多样性：每个等价类中敏感属性值出现的最高频率和最低频率都被约束	
t－近似性模型[15]	所有等价类中敏感属性值的分布与整个数据集中敏感属性值的分布差异不超过阈值 t	链接攻击、同质性攻击、相似性攻击

3.1.2 l－多样性模型 为了抵御同质性攻击，Machanavajjhala A 等[14]提出 l－多样性模型进一步加强对数据集中敏感属性的保护。该模型要求敏感属性在同一个等价类中的值具有多样性，具体可以细分为 4 类，见前文表 1。与 l－多样性模型原理类似的隐私模型还有 p－Sensitive k－匿名模型[16]和（α，k）－匿名模型[17]等。l－多样性相关隐私模型能够抵御链接攻击和同质性攻击，但仍存在相似性攻击风险，例如在匿名数据集的某个等价类中，多个患者患有类似疾病，攻击者便可以通过语义或敏感程度推断出患者疾病。

3.1.3 t－近似性模型 Li N[15]提出 t－近似性模型对等价类中敏感属性值的分布提出限制，以抵御相似性攻击。但该模型明显降低了匿名化数据集质量，尤其是当数据集规模较小或者 k－匿名模型的 k 值较小时数据质量下降严重，只能通过提高阈值 t

来提高输出数据集质量。

3.2 个性化隐私模型

3.2.1 个性化隐私模型分类 个性化隐私模型在传统隐私模型的基础上对不同敏感属性值赋予个性化权重，对高敏感性的敏感属性值进行重点保护。个性化隐私模型可分为两类。一类是仅根据保护需求将敏感值划分不同的保护级别，如（p^+，α）－敏感 k－匿名模型[18]。该模型要求先将敏感属性值依照敏感程度划分不同级别，然后保证在一个等价类中至少包含 p 个不同类别的敏感属性值，避免了同类敏感属性值在单个等价类中的集中分布，能有效实现个性化隐私保护。另一类是通过构建敏感属性泛化结构树，用泛化值取代敏感值实现匿名保护，如个性化（p，k）匿名隐私保护模型[19]。该模型首先对不同敏感值进行评估，然后构建泛化结构树，根据评估分值进行泛化。该方法能有效保护高敏感性的敏感属性值，但也会造成部分敏感属性值丢失，导致数据质量下降。

3.2.2 个性化隐私模型相关研究 近年来，随着个性化数据共享场景增加，个性化隐私模型相关研究也逐渐增多。李文等[20]于 2017 年提出面向医疗数据共享的个性化 l－多样性匿名隐私保护模型，不仅要求匿名化数据满足 Entropy l－多样性模型，而且还将敏感属性值区分为强敏感属性值和弱敏感属性值，限制强敏感属性值出现频率，实现移动医疗系统用户隐私数据保护。2022 年冷建宇[21]针对医疗数据中疾病属性具有双重语义信息的特点，提出个性化的（w，k，d）－匿名模型。该模型不仅按照疾病严重程度进行分级，而且还利用疾病语义层次结构度量不同疾病之间的距离用于约束等价类，从而实现个性化保护。

4 基于聚类的匿名化算法

4.1 实现传统隐私模型的匿名化算法

4.1.1 实现 k－匿名模型算法 作为最基础的模型，实现 k－匿名模型的基于聚类的匿名化算法种类十分丰富，包括 k－means 算法[22]、k－member 算法[23]、平均矢量最大距离算法（maximum distance to average vector algorithm，MDAV）[24]、单程 k－均值算法（one－pass k－means algorithm，OKA）[25]等。k－means 算法[22]是实现 k－匿名模型最简单的算法，通过随机选取聚类中心，多次迭代生成等价类后实现匿名化。k－member 算法[23]和 MDAV 算法[24]原理相似，聚类过程都是逐个元组逐簇进行的，当聚类簇大小达到 k 以后，才开始进行下一个簇的聚类，因此算法性能较差，都具有 $O(n^2)$的时间复杂度。为了降低时间复杂度，Lin J L[25]提出 OKA 算法，通过一次同时随机生成 k 个聚类中心，将聚类过程的时间复杂度降低到了 $O(n^2/k)$。

4.1.2 实现 l－多样性模型的算法 许多实现 l－多样性模型的算法都是在 k－匿名模型算法基础上进行改进的。例如郑珂等[26]基于 k－means 算法提出通过将敏感属性转化为多维向量，然后根据向量距离进行聚类的基于多敏感属性 k－means 算法，能够抵御链接攻击和同质性攻击；夏赞珠等[27]提出基于 MDAV 改进的（k，e）－MDAV 聚类算法，设置敏感属性取值差异参数 e，要求在聚类过程中保证每个簇大小达到 k 且敏感属性取值差异也达到 e 以上，实现抵御同质性攻击的敏感属性保护。Gui Q 等[28]提出基于泛化数据的模糊 C 均值聚类（fuzzy C－means clustering with generalization data，FCMGD）算法，在该算法中，每个元组不是仅分配到单个聚类簇中，而是通过构建隶属度矩阵允许元组对每个聚类都有一个隶属度，然后根据隶属度矩阵调整聚类结果实现 l－多样性模型。

4.1.3 实现 t－近似性模型的算法 为了保证匿名化数据中等价类敏感属性值分布能够与整个数据集分布相同，通常需要首先将整个数据集根据相似敏感属性进行划分，然后再进行聚类。Cao J 等[29]指出敏感属性分类和重分配（sensitive attribute bucketization and redistribution，SABRE）算法框架，首先将原始数据根据敏感属性值的相似性划分为多个组，在构建等价类簇时纳入从各组中等比例选取的元组，以保证生成的等价类敏感属性值与整体敏感属性值的分布趋同，从而实现 t－近似性模型。

Soria – Comas J 等[30]提出可以通过先对敏感属性排序再聚类或先聚类再检查聚类簇是否满足模型要求两种方案实现 t – 近似性模型。Fang Y 等[31]引入完全不相交投影（complete disjoint projections，CO-DIP）方法，用一个单值属性替换每个多值敏感属性，并根据其关联将所有敏感属性分割为一些不相交的子集，然后再分别处理每个子集以满足敏感属性的分布要求。Wang R 等[32]在模糊 C 均值聚类算法基础上，对不满足 t – 近似性模型的聚类簇通过元组抽取再分配的方法实现多敏感属性的 t – 近似性模型。

4.2 实现个性化隐私模型的匿名化算法

实现个性化隐私模型的基于聚类的匿名化算法为了保证敏感属性在各级别的分布，通常需要将整个数据集元组的相似敏感属性进行划分后再进行聚类。如王平水[33]提出的个性化（l，c）– 匿名算法，首先对各敏感属性值的敏感度进行定义，根据敏感属性值的敏感度降序排列构建哈希桶，然后从中选取元组进行聚类使信息损失最小，以保证敏感程度高的属性值得到更高程度保护。对敏感属性进行泛化的个性化匿名模型的聚类算法，如贾俊杰等[19]提出的个性化（p，k）– 匿名隐私保护算法，只需要在普通聚类算法基础上，根据对敏感属性保护的需求，对高敏感性的敏感属性值进行泛化，直至满足使用者需求，便能实现个性化保护。近年来，还出现许多结合敏感属性特点改进的算法。如黄玉蕾等[34]提出的基于多敏感值的个性化隐私保护算法、朱理奥[35]提出的个性化（w，l，k）– 匿名模型等。

4.3 基于聚类的匿名化算法比较分析（表 2）

表 2 代表性基于聚类的匿名化算法特点

聚类算法	隐私模型	时间复杂度	算法优点	不足之处
k – means[22]	k – 匿名模型	O（kn）	实现简单，可扩展性较强	随机选择聚类中心导致聚类效果较差
k – member[23]、MDAV[24]	k – 匿名模型	O（n^2）	逐个簇进行聚类，聚类效果较好	时间复杂度较高；受异常值影响较大
OKA[25]	k – 匿名模型	O（n^2/k）	多个聚类中心同时聚类，控制簇大小，较高效率实现聚类，时间复杂度和信息损失均较低	聚类过程迭代次数较少，受异常值影响较大
V – MDAV[36]	k – 匿名模型	O（n^2）	建立距离矩阵提高效率，聚类簇大小具有弹性，提高聚类效果	时间复杂度较高；受异常值影响较大
（k，e）– MDAV[27]	l – 多样性模型	O（n^2）	在 MDAV 算法基础上对敏感属性值进行约束	时间复杂度较高
FCMGD[28]	l – 多样性模型	O（en^2）	使用模糊聚类方法，有效减少信息损失	时间复杂度较高
个性化（l，c）– 匿名算法[33]	个性化匿名模型	O（n^2）	不仅实现根据敏感程度聚类的个性化匿名模型，同时也满足 t – 近似性模型要求	元组聚类限制过多，信息损失较多

4.3.1 距离度量 距离度量方法不同会影响聚类效果，但许多聚类算法并未给出度量两个元组之间距离的具体方法。通常距离度量与数据属性分类有关。有研究[37]仅提及连续型数据、二元数据等的距离度量方式，未考虑多分类类型数据的距离度量。另有研究[23]提出一种构建分类型数据泛化树的方法，通过比较最小共同父类度量两个多分类类型数据值的距离，以更准确地表示两个元组之间的距离。该方法可以作为改进手段应用于所有聚类算法中。

4.3.2 时间复杂度 从前文表 2 中可以看出，各算法的时间复杂度大小从 O（kn）到 O（en^2）不等。时间复杂度高低与元组在聚类过程中的比较次数有关。时间复杂度低的算法由于元组之间比较次数较少，聚类效果较差，匿名化过程引起的信息损失较多。实现 l – 多样性模型的算法时间复杂度达到 O（n^2），这与约束敏感属性过程中聚类中心需

要与所有元组都进行距离比较有关。

4.3.3 优点及不足 基于聚类的匿名化算法的优点及不足主要受到聚类过程影响，包括聚类中心的选择、聚类簇纳入元组的方式以及等价类的大小等。许多算法是基于原有算法进行改进产生的，例如 V－MDAV 算法在 MDAV 算法的基础上允许每个等价类大小不固定，从而提高簇内元组相似性，减少泛化过程信息损失。

5 基于聚类的匿名化算法应用于医学数据共享隐私保护的建议

5.1 合理选择基于聚类的匿名化算法类型

在医学数据需要共享时，首先需要对共享数据进行分析。如果该数据结构化程度较高，共享时对数据质量具有较高要求，且对匿名化处理时间成本要求较低，那么基于聚类的匿名化算法是比其他匿名化算法更优的选择。选择算法类型时，需要判断不同敏感属性值是否存在不同保护需求，并基于此选择实现传统或个性化的隐私模型算法。在医学数据中往往存在许多需要进行特殊保护的敏感属性值，应当选择实现个性化隐私模型的算法。同时，对敏感属性的保护需求程度也是选择模型的重要依据。从 l－多样性模型到 t－近似性模型等，对敏感属性的分布要求越来越严格，生成的匿名化数据质量也越来越低，因此选择模型算法时需要在加强隐私保护和保证数据质量之间进行权衡。最后，数据集的基本特点也是算法选择的重要影响因素。例如数据集中离群值较多时，不应选择受离群值影响较大的 MDAV 等算法；而数据规模较大或处理设备性能较差，需要在较短时间内获得匿名化结果时，不应选择 FCMGD 等时间复杂度较高的聚类算法。

5.2 灵活改进基于聚类的匿名化算法模型

由于真实世界的数据共享场景千变万化，很难有完全满足使用要求的基于聚类的匿名化算法可供直接使用。因此实际使用时，可以根据数据集特点等对算法进行改进，例如在医学数据共享过程中，如果选择实现 l－多样性模型的（k，e）－MDAV 算法，但数据集中的离群值较多导致聚类效果不够理想时，可以考虑参考加权 k－member 聚类算法进行改进，减少离群值影响。同时，医学数据中通常存在许多缺失值，而大多数基于聚类的匿名化算法都没有讨论存在缺失值时的处理方法。此时则可以参考面向不完整医疗数据集的匿名化聚类算法对缺失值的处理方法[9]，对所选择算法进行改进。此外，在不同场景中衡量匿名化数据集效用的指标不同，可以针对具体方面的效用对算法进行调整改进。例如对面向机器学习用途的数据共享，需要保证匿名化数据的机器学习结果与原始数据的结果相似，可以在匿名化处理过程结合非均衡熵模型，使匿名化数据集具有较好的分类模型训练能力[38]。最后，还可以融合多种算法的优点对所选择的算法进行改进，例如个性化聚类算法与传统聚类算法的融合等。

5.3 加大基于聚类的匿名化算法工具研发力度

目前基于聚类的匿名化算法主要是研究者利用 Java、Python 等编程语言根据算法原理编写程序实现的，实现成本较高。虽然近年来涌现出 sdcMicro 工具包等集合多种基于聚类的匿名化算法工具，但这些工具支持的算法数量均较少且灵活性较差，基于聚类的匿名化算法工具的研发存在大量空白。基于聚类的匿名化算法工具研发一方面可以使一些常见的需要不断重复使用匿名化算法的医学数据共享场景，如基于科研目的的电子病历数据共享等，能够实现快速匿名化处理。这不但可以减少匿名化成本，而且可以提高数据共享积极性，有效保障共享数据隐私安全。另一方面有利于实现数据共享匿名化过程规范化，建立科学统一匿名化要求标准，保障匿名化结果具有相对稳定性，从而提高匿名化结果可靠性，为匿名化评估提供依据。

6 结语

近年来出现的各类传统算法的改进算法模型和个性化隐私模型的匿名化算法在医学领域被广泛应用，研究者在使用这些算法时应尤其注意选择最合

适类型。此外，研究者和医学数据共享者还应当关注数据本身特点和共享目标选择匿名化处理方式，从而实现平衡数据的安全性和可用性。

参考文献

1 吉萍，祝丹娜，谢杨晓虹，等. 健康医疗数据的科研共享应用思考［J］. 医学与哲学，2022，43（1）：5－8.

2 沈洪兵. 大数据时代的临床医学研究——机遇和挑战［J］. 南京医科大学学报（自然科学版），2020，40（3）：303－305.

3 张喆，杨松，王宁，等. 关于新冠肺炎疫情相关数据集成共享平台研究［J］. 统计理论与实践，2020（1）：45－52.

4 辛雨. 科学家呼吁全面开放共享新冠病毒基因组数据［N］. 中国科学报，2021－02－05（1）.

5 崔宇红，王飒. 新型冠状病毒突发公共卫生事件中的数据共享机制研究［J］. 图书情报工作，2020，64（15）：104－111.

6 康盼红. 后疫情时代新冠肺炎档案数据共享平台建设［J］. 档案天地，2021（6）：36－40.

7 高岩，苏东艳. “新冠”疫情下科学数据统筹管理与开放共享的思考［J］. 江苏科技信息，2020，37（10）：8－11.

8 邱春艳，陈可睿. 国内外新冠肺炎数据共享现状分析［J］. 数字图书馆论坛，2022（5）：60－65.

9 裴孟丽. 基于 l－多样性面向缺失医疗数据的匿名算法研究［D］. 郑州：郑州大学，2019.

10 荆学士. 云环境下健康大数据隐私保护技术研究与实现［D］. 成都：电子科技大学，2017.

11 曹惠瑞. 医疗数据发布共享中的隐私保护研究［D］. 石家庄：石家庄铁道大学，2020.

12 SAMARATI P, SWEENEY L. Generalizing data to provide anonymity when disclosing information［C］. Austin: Seventeenth ACM SIGACT－SIGMOD－SIGART Symposium on Principles of Database Systems, 1998.

13 丁有伟，王鹏，胡孔法，等. 一种面向中医药临床数据发布的隐私保护算法［J］. 世界科学技术－中医药现代化，2021，23（7）：2393－2401.

14 MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. L－diversity: privacy beyond k－anonymity［J］. ACM transactions on knowledge discovery from data, 2007, 1（1）: 3.

15 LI N, LI T, VENKATASUBRAMANIAN S. T－closeness: privacy beyond k－anonymity and l－diversity［C］. Istanbul: The 23rd International Conference on Data Engineering (IEEE), 2007.

16 TRUTA T M, VINAY B. Privacy protection: p－sensitive k－anonymity property［C］. Atlanta: 22nd International Conference on Data Engineering Workshops (ICDEW'06) (IEEE), 2006.

17 WONG C W, LIU Y, YIN J, et al. (α, k)－anonymity based privacy preservation by lossy join［C］. Berlin: Web－Age Information Management, 2007.

18 黄石平，顾金媛. 一种基于（p^+，α）－敏感 k－匿名的增强隐私保护模型［J］. 计算机应用研究，2014，31（11）：3465－3468.

19 贾俊杰，闫国蕾. 一种个性化（p，k）匿名隐私保护算法［J］. 计算机工程，2018，44（1）：176－181.

20 李文，黄丽韶，罗恩韬. 移动医疗中个性化 l－多样性匿名隐私保护模型［J］. 计算机科学与探索，2018，12（5）：761－768.

21 冷建宇. 面向医疗信息隐私保护的匿名化技术研究［D］. 南京：南京邮电大学，2021.

22 VAIDYA J, CLIFTON C. Privacy－preserving k－means clustering over vertically partitioned data［C］. Washington, DC: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

23 BYUN J W, KAMRA A, BERTINO E, et al. Efficient k－anonymization using clustering techniques［C］. Berlin: International Conference on Database Systems for Advanced Applications, 2007.

24 DOMINGO－FERRER J, MATEO－SANZ J M. Practical data－oriented microaggregation for statistical disclosure control［J］. IEEE transactions on knowledge & data engineering, 2002, 14（1）: 189－201.

25 LIN J L, WEI M C. An efficient clustering method for k－anonymization［C］. Chung－Li: International Workshop on Privacy and Anonymity in Information Society (PAIS), 2008.

26 郑珂. 基于 Sensitive－k 匿名模型的隐私保护算法研究［D］. 哈尔滨：哈尔滨工程大学，2021.

27 夏赞珠，韩建民，于娟，等. 用于实现（k，e）－匿名模型的 MDAV 算法［J］. 计算机工程，2010（15）：165－167.

28 GUI Q, LV Y, CHENG X, et al. Data anonymous method based on fuzzy clustering［C］. Guilin: 2019 4th International Conference on Intelligent Information Processing, 2019.

（下转第 78 页）

tive cardiology, 2021, 28 (15): 1682 - 1690.
3 TAYLOR R S, LONG L, MORDI I R, et al. Exercise - based rehabilitation for heart failure: Cochrane systematic review, meta - analysis, and trial sequential analysis [J]. JACC heart failure, 2019, 7 (8): 691 - 705.
4 PELLICCIA A, SHARMA S, GATI S, et al. 2020 ESC guidelines on sports cardiology and exercise in patients with cardiovascular disease [J]. European heart journal, 2021, 42 (1): 17 - 96.
5 SUTTON R T, PINCOCK D, BAUMGART D C, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success [J]. NPJ digital medicine, 2020, 3 (1): 17.
6 SHEN B. Translational informatics: sports and exercise medicine [M]. Singapore: Springer, 2022.
7 SHEN L, BAI J, WANG J, et al. The fourth scientific discovery paradigm for precision medicine and healthcare: challenges ahead [J]. Precision clinical medicine, 2021, 4 (2): 80 - 84.
8 HANSEN D, DENDALE P, CONINX K, et al. The European Association of Preventive Cardiology exercise prescription in everyday practice and rehabilitative training (EXPERT) tool: a digital training and decision support system for optimized exercise prescription in cardiovascular disease. Concept, definitions and construction methodology [J]. European journal of preventive cardiology, 2020, 24 (10): 1017 - 1031.
9 PESCATELLO L S, WU Y, PANZA G A, et al. Development of a novel clinical decision support system for exercise prescription among patients with multiple cardiovascular disease risk factors [J]. Mayo clinic proceedings innovations, quality & outcomes, 2020, 5 (1): 193 - 203.
10 ZHANG K, BAO T, WU R, et al. PAHFKB: a knowledge base and an online service for personalized physical activity in prevention and intervention of heart failure [EB/OL]. [2022 - 07 - 10]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4147835.
11 谭玲, 鄂海红, 匡泽民, 等. 医学知识图谱构建关键技术及研究进展 [J]. 大数据, 2021, 7 (4): 80 - 104.
12 秦川, 祝恒书, 庄福振, 等. 基于知识图谱的推荐系统研究综述 [J]. 中国科学: 信息科学, 2020, 50 (7): 937 - 956.
13 CHEN H, SULTAN S, TIAN Y, et al. Fast and accurate network embeddings via very sparse random projection [EB/OL]. [2022 - 08 - 21]. https://dl.acm.org/doi/10.1145/3357384.3357879.
14 芦威. 推荐结果多样性的评价方法及优化算法研究 [D]. 北京: 北京交通大学, 2019.

(上接第 71 页)

29 CAO J, KARRAS P, KALNIS P, et al. SABRE: a sensitive attribute bucketization and redistribution framework for t - closeness [J]. The VLDB journal, 2011, 20 (1): 59 - 81.
30 SORIA - COMAS J, DOMINGO - FERRER J, SANCHEZ D, et al. T - closeness through microaggregation: strict privacy with enhanced utility preservation [J]. IEEE transactions on knowledge and data engineering, 2015, 27 (11): 3098 - 3110.
31 FANG Y, ASHRAFI M Z, NG S K. Privacy beyond single sensitive attribute [C]. Berlin: International Conference on Database and Expert Systems Applications, 2011.
32 WANG R, ZHU Y, CHEN T S, et al. Privacy - preserving algorithms for multiple sensitive attributes satisfying t - closeness [J]. Journal of computer science and technology, 2018, 33 (6): 1231 - 1242.
33 王平水, 王建东. 一种基于聚类的个性化 (l, c) - 匿名算法 [J]. 计算机工程与应用, 2012, 48 (23): 5.
34 黄玉蕾, 林青, 戴慧珺. 基于多敏感值的个性化隐私保护算法 [J]. 计算机与数字工程, 2016, 44 (9): 1761 - 1765, 1800.
35 朱理奥, 曹天杰. 量化敏感度的个性化数据发布模型 [J]. 计算机应用与软件, 2022 (8): 39.
36 SOLANAS A, MARTINEZ - BALLESTE A, DOMINGO - FERRER J. V - MDAV: a multivariate microaggregation with variable group size [C]. Rome: 17th COMPSTAT Symposium of the IASC, 2006.
37 徐瑞. 边缘计算中的数据隐私保护技术研究 [D]. 北京: 北京交通大学, 2020.
38 唐明坤, 钱庆, 张丽鑫, 等. 生物医学数据匿名化工具 ARX 研究及启示 [J]. 中华医学图书情报杂志, 2022, 31 (2): 19 - 29.