



(12) 发明专利申请

(10) 申请公布号 CN 112364914 A

(43) 申请公布日 2021.02.12

(21) 申请号 202011245565.8

(22) 申请日 2020.11.10

(71) 申请人 郑州大学

地址 450001 河南省郑州市高新技术开发区科学大道100号

(72) 发明人 叶阳东 徐富国 胡世哲

(51) Int.Cl.

G06K 9/62 (2006.01)

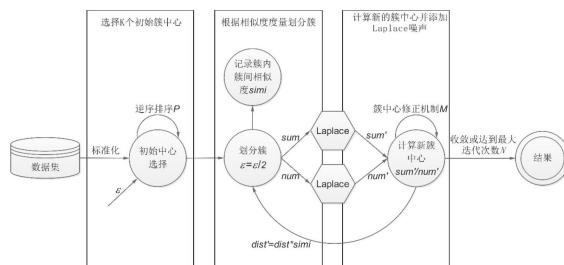
权利要求书2页 说明书5页 附图3页

(54) 发明名称

基于簇相似度与变换不变性的差分隐私k均值聚类方法

(57) 摘要

本发明公开了一种基于簇相似度与变换不变性的差分隐私k均值聚类方法，包括：对数据集进行预处理；对样本被选中作为下一个初始簇中心的概率P进行逆序排序，并使用轮盘法选取K个初始簇中心；计算每个样本到K个簇中心的欧式距离和到每个簇集合的相似度的乘积，并将其作为相似度度量指标；根据相似度度量指标划分簇；计算新的簇中心 u'_j ，并使用差分隐私的Laplace噪声机制对新的簇中心进行隐私保护；根据差分隐私的变换不变性，对Laplace噪声机制扰动后的新的簇中心进行修正；重复步骤直到满足收敛条件或达到最大迭代次数N；本发明解决了现有差分隐私聚类方法在添加Laplace噪声扰动后其结果可用性较差的问题。



1. 基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,按照以下步骤实施:

步骤1、对数据集进行预处理;

步骤2、给定隐私保护预算 ϵ ,随机选取一个样本作为首个初始簇中心;

步骤3、计算其他样本被选中作为下一个初始簇中心的概率P,并对此概率P进行逆序排序;

步骤4、用轮盘法选取下一个初始簇中心,循环执行步骤3和4直至获取K个初始簇中心;

步骤5、计算每个样本到K个簇中心的欧式距离和到每个簇集合的相似度,并将它们的乘积作为划分簇时的相似度度量指标;

步骤6、根据相似度度量指标划分簇;

步骤7、计算新的簇中心 u'_j ,使用差分隐私的Laplace噪声机制对新的簇中心进行隐私保护;

步骤8、根据差分隐私的变换不变性,对Laplace噪声机制扰动后的新的簇中心进行修正;

步骤9、满足收敛条件或达到最大迭代次数N则聚类结束,输出结果,否则返回步骤5。

2. 按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤1中预处理是指将所有样本归一化到 $[0,1]^d$ 空间内,d表示数据集中样本的维度。

3. 按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤3具体为:计算其他样本到当前已有簇中心之间的最短距离,即与最近簇中心的欧式距离,并计算这些最短距离之和。将所述最短距离与最短距离之和相除得到其他样本被选中作为下一个初始簇中心的概率,并对此概率进行逆序排序。

4. 按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤5中,用D表示聚类所用数据集, x_i 表示数据集的一个样本, C_j 表示某一个簇集合,C表示所有簇集合,其中 $1 \leq i \leq |D|, 1 \leq j \leq |C|$ 。每个样本到每个簇集合的相似度计算方法为:

$$\text{similarity}(x_i, C_j) = \frac{1}{|C_j|} \sum_{y \in C_j} \|x_i - y\|_2^2, C_j \in C,$$

如果样本 x_i 属于簇集合 C_j ,则称 $\text{similarity}(x_i, C_j)$ 为簇内相似度,否则称为簇间相似度。用 $\text{dist}(x_i, u_j)$ 表示样本 x_i 与簇中心 u_j 的欧式距离,则相似度度量指标计算方法为:

$$\text{dist}'(x_i, u_j) = \text{dist}(x_i, u_j) * \text{similarity}(x_i, C_j).$$

5. 按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤6具体为:比较划分簇时的相似度度量指标的大小,选取相似度度量指标值最小的所属簇作为划分簇的结果。

6. 按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤7具体为:计算每个簇集合的属性向量之和sum和样本数量num;根据数据集特征,得到簇中样本属性向量之和sum的全局敏感度 $\Delta f_{sum} = d$,样本数量num的全局敏感度 $\Delta f_{num} = 1$,进而得到整个数据集的查询敏感度 $\Delta f = d + 1$;将隐私保护预算缩小为原来的

一半 $\epsilon = \epsilon/2$;计算第t次迭代时添加的Laplace噪声 $noise^t = Lap(\Delta f * 2^t / \epsilon)$;分别对属性向量之和sum和样本数量num添加Laplace噪声,即 $sum' = sum + noise^t$, $num' = num + noise^t$;计算新的簇中心 $u'_j = sum' / num'$ 。

7.按照权利要求1所述的基于簇相似度与变换不变性的差分隐私k均值聚类方法,其特征在于,所述步骤8具体为:根据差分隐私的变换不变性,选择一种随机性独立于原始数据集和Laplace分布的簇中心修正机制M,并将其作用到步骤7中新的簇中心 u'_j 上,使得其属性向量均变换到归一化范围 $[0,1]^d$ 空间内,计算方法为:

$$M\left(u'_j = \frac{sum'}{num'}\right) \in [0,1]^d.$$

基于簇相似度与变换不变性的差分隐私k均值聚类方法

技术领域

[0001] 本发明属于隐私保护技术领域,尤其涉及一种基于簇相似度与变换不变性的差分隐私k均值聚类方法。

背景技术

[0002] K均值聚类作为一种典型的无监督数据挖掘方法,可以从海量数据中挖掘未知知识和潜在价值,该方法已经广泛应用于刻画群体特征、市场预测和广告推送等多个领域。然而在挖掘有用信息的同时,可能会泄露数据中的个人隐私信息,对用户造成不可估量的威胁和损失。因此在k均值聚类分析过程中,需要保护数据中的个人隐私信息,并保证最终聚类结果的可用性。

[0003] 随着攻击者所拥有的背景知识越来越多,传统的隐私保护技术,如k-anonymity, t-closeness等,已经难以保证数据的安全性。差分隐私作为一种新型的严格数学证明的隐私保护模型而广泛应用于数据发布、数据挖掘和机器学习等多个领域。将差分隐私应用于k均值聚类分析上能够很好地保护源数据,然而现有的差分隐私k均值聚类方法仍存在许多不足,如聚类结果的可用性较差、迭代次数受限、簇中心偏离真实簇中心等,因此如何在提供隐私保护的同时提高k均值聚类结果的可用性,是如今亟需解决的问题。

发明内容

[0004] 本发明实施例提供了一种基于簇相似度与变换不变性的差分隐私k均值聚类方法,旨在解决现有差分隐私聚类方法添加Laplace噪声后其结果可用性较差的问题。

[0005] 为实现上述目的,本发明实施例提供了一种基于簇相似度与变换不变性的差分隐私k均值聚类方法,包括以下步骤:

[0006] 步骤1、对数据集进行预处理;

[0007] 步骤2、给定隐私保护预算 ϵ ,随机选取一个样本作为首个初始簇中心;

[0008] 步骤3、计算其他样本被选中作为下一个初始簇中心的概率P,并对此概率P进行逆序排序;

[0009] 步骤4、用轮盘法选取下一个初始簇中心,循环执行步骤3和4直至获取K个初始簇中心;

[0010] 步骤5、计算每个样本到K个簇中心的欧式距离和到每个簇集合的相似度,并将它们的乘积作为划分簇时的相似度度量指标;

[0011] 步骤6、根据相似度度量指标划分簇;

[0012] 步骤7、计算新的簇中心 u'_j ,使用差分隐私的Laplace噪声机制对新的簇中心进行隐私保护;

[0013] 步骤8、根据差分隐私的变换不变性,对Laplace噪声机制扰动后的新的簇中心进行修正;

[0014] 步骤9、满足收敛条件或达到最大迭代次数N则聚类结束,输出结果,否则返回步骤

5。

[0015] 本发明的特点还在于，

[0016] 步骤1中，预处理是指将所有样本归一化到 $[0,1]^d$ 空间内，d表示数据集中样本的维度。

[0017] 步骤3具体为：计算其他样本到当前已有簇中心之间的最短距离，即与最近簇中心的欧式距离，并计算这些最短距离之和。将所述最短距离与最短距离之和相除得到其他样本被选中作为下一个初始簇中心的概率，并对此概率进行逆序排序。

[0018] 步骤5中，用D表示聚类所用数据集， x_i 表示数据集的一个样本， C_j 表示某一个簇集合，C表示所有簇集合，其中 $1 \leq i \leq |D|$, $1 \leq j \leq |C|$ 。每个样本到每个簇集合的相似度计算方法如公式(1)所示：

$$[0019] \text{similarity}(x_i, C_j) = \frac{1}{|C_j|} \sum_{y \in C_j} \|x_i - y\|_2^2, C_j \in C \quad (1)$$

[0020] 如果样本 x_i 属于簇集合 C_j ，则称 $\text{similarity}(x_i, C_j)$ 为簇内相似度，否则称为簇间相似度。用 $\text{dist}(x_i, u_j)$ 表示样本 x_i 与簇中心 u_j 的欧式距离，则相似度度量指标计算方法如公式(2)所示：

$$[0021] \text{dist}'(x_i, u_j) = \text{dist}(x_i, u_j) * \text{similarity}(x_i, C_j) \quad (2)$$

[0022] 步骤6具体为：比较划分簇时的相似度度量指标的大小，选取相似度度量指标值最小的所属簇作为划分簇的结果。

[0023] 步骤7具体为：计算每个簇集合的属性向量之和sum和样本数量num；根据数据集特征，得到簇中样本属性向量之和sum的全局敏感度 $\Delta f_{sum} = d$ ，样本数量num的全局敏感度 $\Delta f_{num} = 1$ ，进而得到整个数据集的查询敏感度 $\Delta f = d + 1$ ；将隐私保护预算缩小为原来的一半 $\epsilon = \epsilon/2$ ；计算第t次迭代时添加的Laplace噪声 $\text{noise}^t = \text{Lap}(\Delta f * 2^t / \epsilon)$ ；分别对属性向量之和sum和样本数量num添加Laplace噪声，即 $\text{sum}' = \text{sum} + \text{noise}^t$, $\text{num}' = \text{num} + \text{noise}^t$ ；计算新的簇中心 $u'_j = \text{sum}' / \text{num}'$ 。

[0024] 步骤8具体为：根据差分隐私的变换不变性，选择一种随机性独立于原始数据集和Laplace分布的簇中心修正机制M，并将其作用到步骤7中新的簇中心 u'_j 上，使得其属性向量均变换到归一化范围 $[0,1]^d$ 空间内，其计算方法如公式(3)所示：

$$[0025] M(u'_j) = \frac{\text{sum}'}{\text{num}'} \in [0,1]^d \quad (3)$$

[0026] 本发明的有益效果为：

[0027] 一种基于簇相似度与变换不变性的差分隐私k均值聚类方法，首先，针对k均值聚类结果对初始中心选择的敏感问题，考虑到使用轮盘法选择初始簇中心时样本被选中概率的分布特性，将样本被选中作为初始簇中心的概率进行逆序排序，使得较远的点被作为下一个初始簇中心的可能性较大。其次，为降低随机噪声带来的误差影响，重新定义相似度度量指标，将每个样本点的簇内和簇间相似度与传统欧式距离的乘积作为划分簇时的相似度度量指标。最后，由于在聚类迭代过程中簇中心存在偏离的可能性，而且迭代次数越大，簇中心偏离的可能性就会越大，根据差分隐私的变换不变性原理，给定一种随机性独立于原始数据集和Laplace分布的机制M，将簇中心属性向量均修正到 $[0,1]^d$ 空间范围内，以此来解决簇中心偏离问题，但隐私性仍然满足差分隐私。通过与现有方法比较，在同等隐私保护

程度下,本发明具有较高的聚类结果可用性。除此之外,当迭代次数很大时,本发明也能保持较高的聚类结果可用性。

附图说明

- [0028] 图1是本发明实施例提供的一种基于簇相似度与变换不变性的差分隐私k均值聚类方法框架图;
- [0029] 图2是本发明实施例提供的实验数据集;
- [0030] 图3是本发明实施例提供的不同隐私保护预算 ϵ 下“Iris”数据集可用性对比图;
- [0031] 图4是本发明实施例提供的不同隐私保护预算 ϵ 下“Texture”数据集可用性对比图;
- [0032] 图5是本发明实施例提供的不同迭代次数N下“Iris”数据集可用性对比图;
- [0033] 图6是本发明实施例提供的不同迭代次数N下“Texture”数据集可用性对比图。

具体实施方式

[0034] 为使本申请实施例的目的、技术方案及优点更加清楚,下面结合附图和具体实施方式对本发明进一步详细说明,应理解这些实例仅用于说明本发明而不同于限制本发明的范围,在阅读了本发明之后,本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0035] 本发明一种基于簇相似度与变换不变性的差分隐私k均值聚类方法,其整体框架如图1所示,具体按照以下步骤实施:

[0036] 步骤1、对数据集进行预处理,将所有样本归一化到 $[0, 1]^d$ 空间内,d表示数据集中样本的维度;

[0037] 步骤2、给定隐私保护预算 ϵ ,随机选取一个样本作为首个初始簇中心;

[0038] 步骤3、计算其他样本到当前已有簇中心之间的最短距离,即与最近簇中心的欧式距离,并计算这些最短距离之和。将所述最短距离与最短距离之和相除得到其他样本被选中作为下一个初始簇中心的概率P,并对此概率P进行逆序排序;

[0039] 步骤4、用轮盘法选取下一个初始簇中心,循环执行步骤3和4直至获取K个初始簇中心;

[0040] 步骤5、用D表示聚类所用数据集, x_i 表示数据集的一个样本, C_j 表示某一个簇集合,C表示所有簇集合,其中 $1 \leq i \leq |D|, 1 \leq j \leq |C|$ 。计算每个样本到K个簇中心的欧式距离 $dist(x_i, u_j)$ 和到每个簇集合的相似度 $similarity(x_i, C_j) = \frac{1}{|C_j|} \sum_{y \in C_j} \|x_i - y\|_2^2, C_j \in C$,并将它们的乘积作为划分簇时的相似度度量指标,即 $dist'(x_i, u_j) = dist(x_i, u_j) * similarity(x_i, C_j)$,如果样本 x_i 属于簇集合 C_j ,则称 $similarity(x_i, C_j)$ 为簇内相似度,否则称为簇间相似度。对于一个样本来说,簇内相似度一定小于簇间相似度,这就使得其与所属簇集合的相似度度量指标的值较与其他簇集合的相对会变小,也就是相似度相对会变高;

[0041] 步骤6、根据相似度度量指标划分簇,具体为:比较划分簇时的相似度度量指标的大小,选取相似度度量指标值最小的所属簇作为划分簇的结果;

[0042] 步骤7、计算新的簇中心,以差分隐私的Laplace噪声机制对新的簇中心进行隐私

保护,具体为:计算每个簇集合的属性向量之和sum和样本数量num;根据数据集特征,得到簇中样本属性向量之和sum的全局敏感度 $\Delta f_{sum}=d$,样本数量num的全局敏感度 $\Delta f_{num}=1$,进而得到整个数据集的查询敏感度 $\Delta f=d+1$;将隐私保护预算缩小为原来的一半 $\epsilon=\epsilon/2$;计算第t次迭代时添加的Laplace噪声 $noise^t=Lap(\Delta f*2^t/\epsilon)$;计算新的簇中心 $u'_j=(sum+noise^t)/(num+noise^t)$;

[0043] 步骤8、根据差分隐私的变换不变性,对Laplace噪声机制扰动后的簇中心进行修正,具体为:根据差分隐私的变换不变性,选择一种随机性独立于原始数据集和Laplace分布的簇中心修正机制M,并将其作用到步骤7中新的簇中心 u'_j 上,使得其属性向量均变换到

归一化范围 $[0,1]^d$ 空间内,即 $M(u'_j = \frac{sum}{num}) \in [0,1]^d$;

[0044] 步骤9、满足收敛条件或达到最大迭代次数N则聚类结束,输出结果,否则返回步骤5。

[0045] 为验证本发明在同等隐私保护水平和不同迭代次数下具有较高的聚类结果可用性,通过仿真实验给出对比实验。实验环境为Intel (R) Core (TM) i7 CPU@3.20GHZ,16GB内存,Windows10 64位操作系统,使用Java语言实现所有算法。

[0046] 实验所用数据集为UCI Knowledge Discovery Archive database中的“Iris”数据集和“Texture”数据集。这两个数据集均给出了样本的真实分类标签,因此可以评估聚类算法的可用性。

[0047] 表1实验数据集

	数据集	属性数	类型	样本数	分类数
[0048]	Iris	3	real	150	3
	Texture	40	real	5500	11

[0049] 实验工作通过对本发明算法与现有差分隐私k均值算法来验证本发明算法的有效性,其中HADPK-means++代表本发明算法,其他七种代表对比算法。

[0050] F-measure是聚类结果可用性评估的度量方法之一,其将准确率和召回率进行综合,同时考虑两者,值越大,表示聚类结果可用性越好,反之,则结果可用性越差。假设CLURESULT表示真实的划分结果,CLURESULT'表示聚类算法的聚类结果,分别用 U_i 和 V_i 表示CLURESULT和CLURESULT'中的第i个簇,cover_i表示 U_i 和 V_i 中样本交集的数量,则第i个簇的准确率 $P_i = \frac{cover_i}{|V_i|}$,召回率 $R_i = \frac{cover_i}{|U_i|}$,由 P_i 和 R_i 的加权调和平均可得到第i个簇的F-measure,记为 F_i ,则 $F_i = \frac{2 * P_i * R_i}{P_i + R_i}$ 。最后,对每个簇的 F_i 求加权平均,可得到整体聚类结果的

$$\text{可用性度量 } F-measure = \frac{\sum_{i=1}^k (|c_i| * F_i)}{\sum_{i=1}^k |c_i|}.$$

[0051] (1) 不同隐私保护预算 ϵ 对可用性的影响分析

[0052] 为对比不同隐私保护预算 ϵ 下可用性的影响分析,设置隐私保护预算 ϵ 由0.1至1.0,步长为0.1,并设置最大迭代次数为2。同时,对设置的每个隐私保护预算,运行20次得

到F-measure的平均值当作可用性度量指标。

[0053] 由图3和图4的实验结果可以看出,本发明不仅在同等保护水平下优于对比算法,而且在隐私保护水平较高时,仍能提供较好的聚类效果。这是因为本发明通过改进初始中心选择、提出新的划分簇的相似度度量指标和提出簇中心修正机制三种方式降低和修正了簇中心偏离带来的误差影响,所以当隐私保护预算 ϵ 较小,添加的Laplace随机噪声较大时也能保持较高水平的聚类结果可用性。

[0054] (2) 不同迭代次数N对可用性的影响分析

[0055] 现有的差分隐私k均值聚类算法随着迭代次数的增加,簇中心偏离的程度会越来越大,导致聚类结果可用性大大降低,但本发明不受迭代次数的影响,由此增加了不同迭代次数N对可用性影响的分析实验。本次实验设置隐私保护预算 $\epsilon=1.0$,迭代次数N={2,3,4,5,6,7,8,9,10},实验结果如图5和图6所示。可以看出,对比算法对迭代次数均非常敏感,在迭代次数较小时,聚类效果较好,之后随着迭代次数的增加,其可用性大大降低。而本发明当迭代次数增加,噪声相对越来越大时,其可用性不仅优于对比算法,而且在两种数据集上表现均较为稳定,因此具有很强的鲁棒性。这是因为本发明考虑了样本点的簇内和簇间相似度,降低了Laplace随机噪声带来的误差影响,而且在计算新簇中心上,修正了偏离的簇中心属性向量。

[0056] 本发明针对差分隐私k均值聚类方法中,当隐私保护水平较高和迭代次数较大时簇中心偏离的问题,从初始中心选择、样本点划分簇和计算新的簇中心三个方面,改进了轮盘法选择初始中心,更加稳定和准确;使用欧式距离与簇内和簇间相似度来重新定义相似度度量指标,降低了Laplace随机噪声带来的误差影响;使用簇中心修正机制,修正了偏离的簇中心属性向量,从根本上解决了簇中心偏离导致的所有样本点划分到同一个簇中的问题。两组实验结果表明,本发明不仅没有局限于迭代次数的限制,而且在较高隐私保护程度下,聚类结果的可用性仍能保持较高水平。

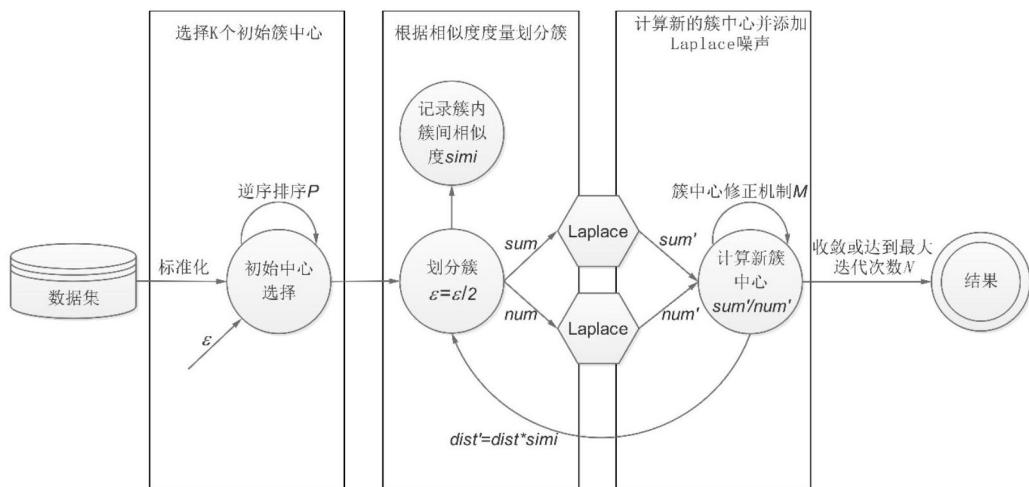


图1

数据集 [*]	属性数 [*]	类型 [*]	样本数 [*]	分类数 [*]
Iris [*]	3 [*]	real [*]	150 [*]	3 [*]
Texture [*]	40 [*]	real [*]	5500 [*]	11 [*]

图2

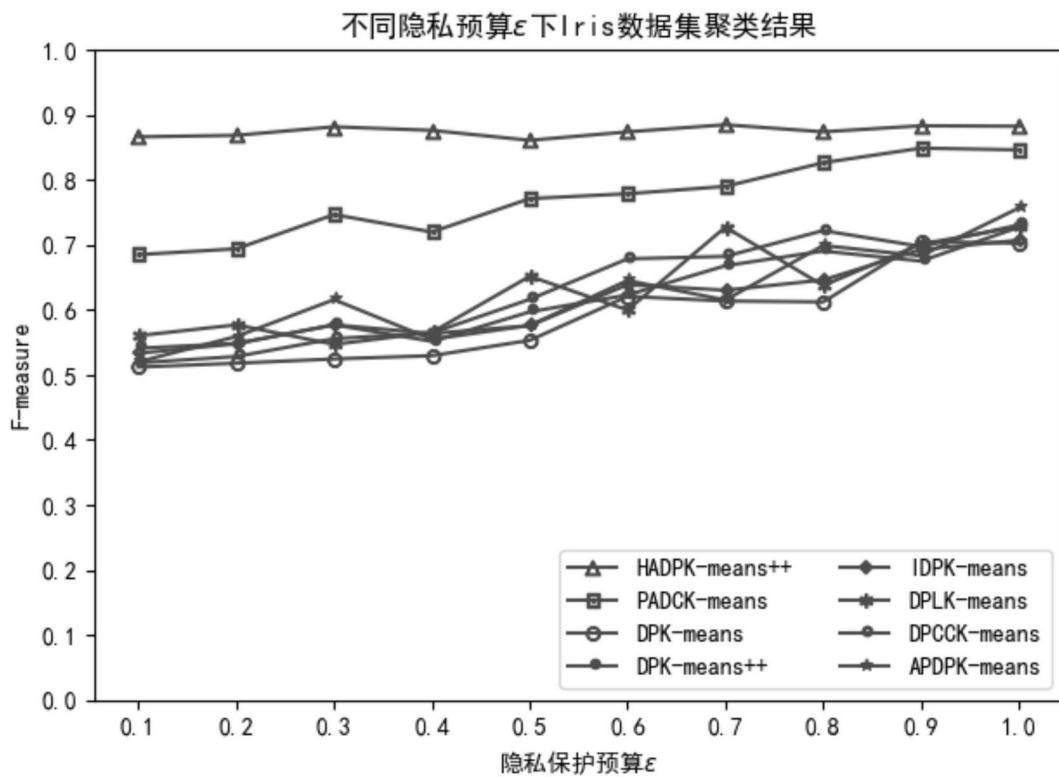


图3

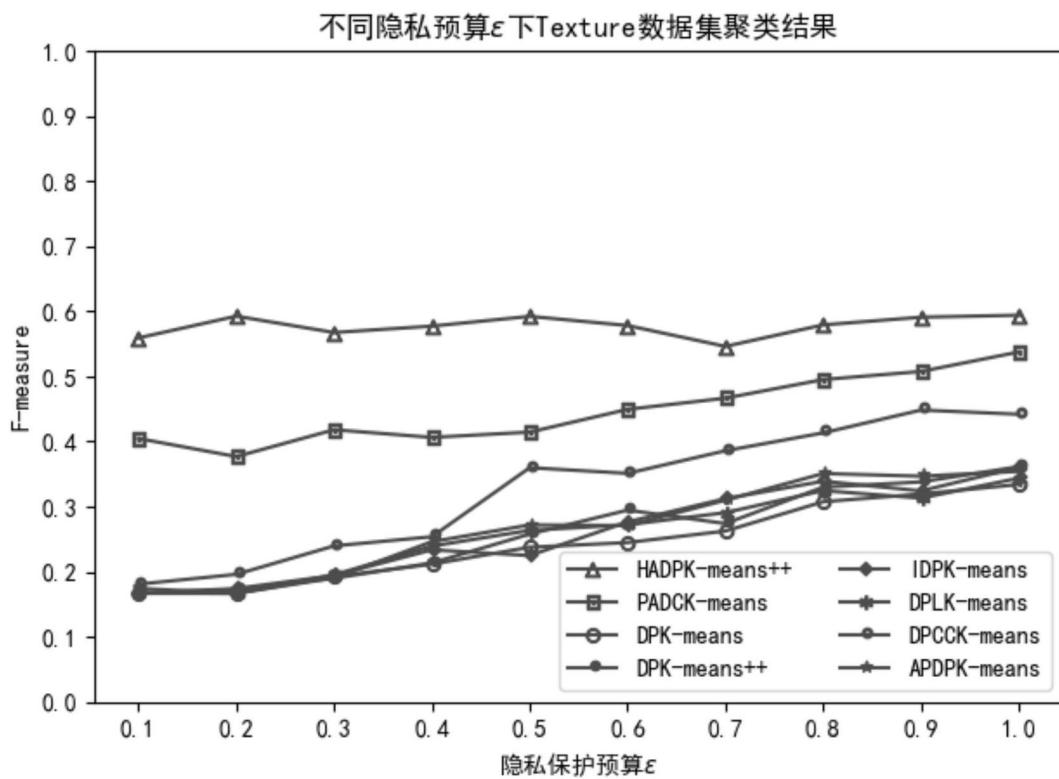


图4

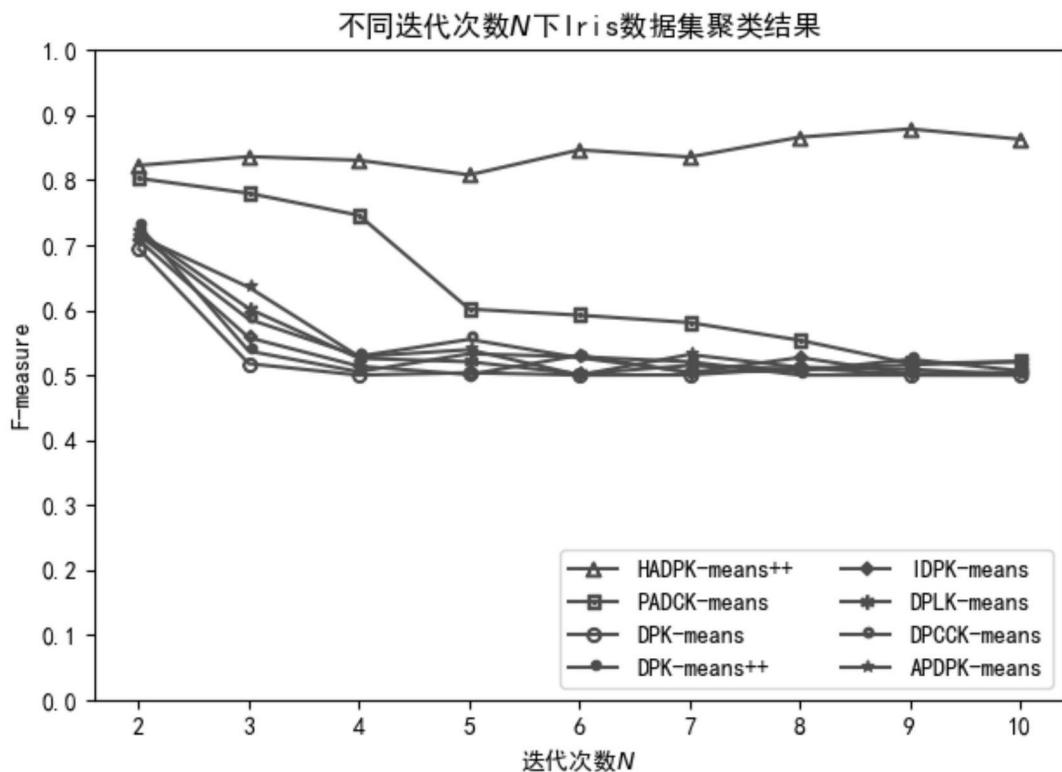


图5

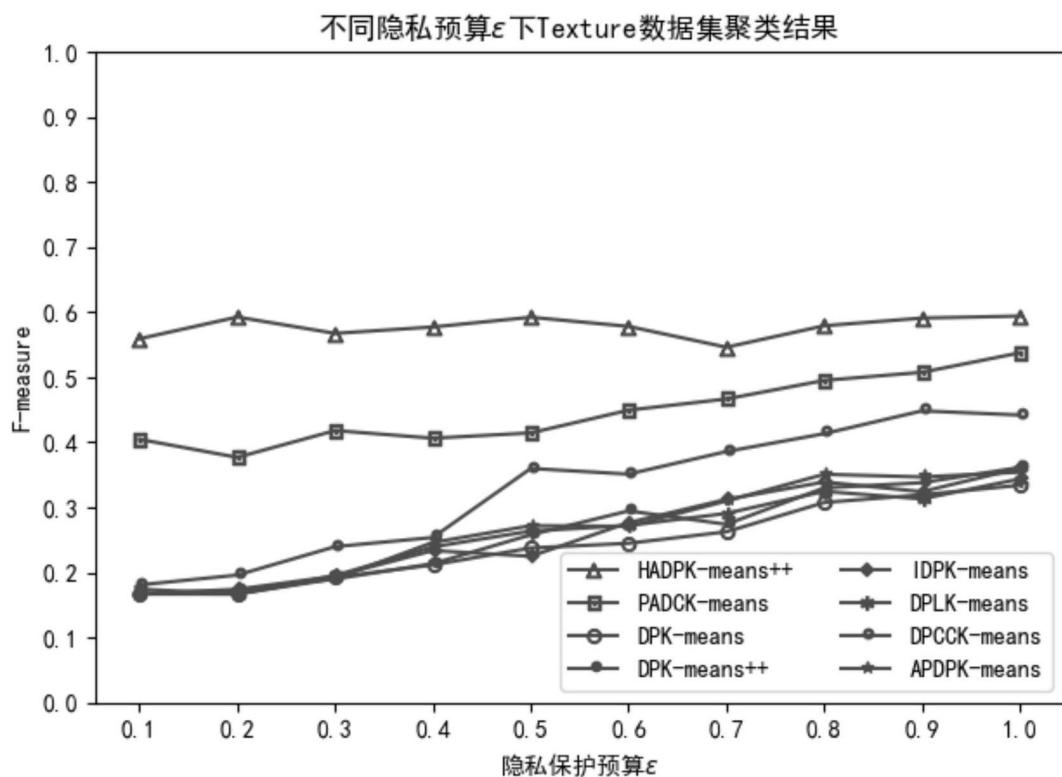


图6