一种改进的面向差分隐私保护的k-means聚类算法

摘要

**首先介绍了差分隐私保护的研究背景和差分隐私保护的基本原理和方法。分析了k-means算法的隐私泄露问题，针对传统面向差分隐私保护k-means算法，存在对初始中心点的选取具有一定盲目性的原因而导致聚类结果的可用性较低的问题，文章提出一种采用指数加噪机制与密度估计相结合的方法，来选取初始聚类中心，从而保证了初始中心挑选的合理性，同时也保障样本数据的隐私性，实验结果表明，新提出的方法可以使聚类结果的可用性显著提高。**

**关键词： 隐私保护; 差分隐私; k-means；**[**聚类算法**](http://kns.cnki.net/kcms/detail/knetsearch.aspx?dbcode=CJFD&sfield=kw&skey=%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95&code=&uid=WEEvREcwSlJHSldRa1FhcTdWa2JKYS9OblhIM0taRTJsa3Nlemc2Y2hsMD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoHbKxJw!!)

1. 引言

**随着大数据时代的来临[1]，大数据的应用已经渗透到医疗、金融、保险、交通、社交等各个领域，人们通过数据的采集、分析和处理获取数据中蕴含的价值，然而，在大数据开发应用过程中，用户隐私与敏感信息极易遭到泄露，给用户带来风险和危害。面对这种情况，通常会对数据进行匿名处理，即使如此，在大数据环境下，因为数据的动态性和大数据平台的开放性，使得攻击者很容易掌握大量攻击背景知识，不用直接获取数据集中的目标信息，而通过背景知识间接分析出目标信息，达到窃取用户的隐私信息的目的。在这样的背景下，Dwork 提出了差分隐私（differential privacy，DP） [2] 的概念。差分隐私保护通过对目标数据添加噪声的方式，在数据集中任意添加或者删除一条记录不会影响最终的查询结果，这样即便攻击者拥有足够多背景知识（即攻击者已知除目标记录外的所有记录的信息），也无法从最终的输出中推测目标信息，从而更好地达到保证隐私保护目的。[3]**

1. 差分隐私保护

**差分隐私[4]是基于数据失真保护技术, 通过添加随机噪声达到数据失真的效果, 同时保持某些数据的属性或者数据在统计方面的性质不变。差分隐私保护技术既能保证数据失真，同时也不会对数据集的可用性产生较大的影响。**

**定义[5] [1]：设有随机算法 M，PM为M所有可能的输出构成的集合．对于任意两个邻近数据集D和以及PM的任何子集SM，若算法 M 满足**

**Pr[M(D) ∈ S M ]≤eε Pr[M(D’) ∈ S M ]（1）**

**则称算法 M 提供 ε- 差分隐私保护，其中参数 ε 称为隐私保护预算。**

1. k-means聚类算法中的隐私泄露

**k-means聚类算法本质是基于距离的聚类分析, 其核心思想是找到k个簇中心点，然后将数据集中的各数据点划分到聚类中心，使得数据点与聚类中心的距离最小，这是一个不断迭代更新中心点的过程,具体步骤为:**

**step1：从数据集中随机选取k个样本作为初始聚类中心C={c1，c2…,Ck}；**

**step2：针对数据集中的每个样本xi,计算它到k个聚类中心的距离并将其分到距离其最小的聚类中心所对应的类中；**

**step3：针对每个类别ci,重新计算它的聚类中心Ci=;**

**step4:重复步骤2和步骤3直到聚类中心不再变化。**

**在聚类过程很有可能会泄露隐私，一是因为step2中计算数据集中的样本与聚类中心距离时，若攻击者获知每次迭代过程中某一数据样本与各聚类中心的距离，则可以通过这些距离推测出数据样本的属性值，而且迭代次数越多，推测出的可能性越大，隐私泄露风险越大；二是因为在聚类完成后，算法收敛，得到最终聚类中心点，最终将发布聚类中心，根据聚类中心点，攻击者结合背景知识，同样可以推断出数据样本的属性值，所以为防止隐私泄露，可以通过添加噪声的方式，完成面向差分隐私k-means算法的设计。**

1. 差分隐私k-means聚类算法

４.１差分隐私k-means聚类方法思想

**通过分析可知，k-means算法[4]隐私泄露的关键是发布聚类中心，为防止这一过程造成的隐私泄露，因此，可以在聚类过程中给聚类中心点，添加合适噪声，使得隐私泄露的风险满足差分隐私保护的要求，算法具体步骤为：**

**step1：从数据集中随机选取k个样本并添加噪声，为初始聚类中心c={c1，c2…,Ck}；**

**step2：针对数据集中的每个样本xi,计算它到k个聚类中心的距离并将其分到距离其最小的聚类中心所对应的类中，得到k个点的集合C={C1，C2…,Ck}；；**

**step3：针对每个集合Ci,计算各个点之和sum=，和C集合中点数，并对sum与，并以此为新的中心点。**

**step4:重复步骤2和步骤3直到聚类中心不再变化。**

**该算法对每次迭代的聚类中心添加噪声，使中心点具有一定偏差，且迭代次数越多，偏差越大，最终的聚类结果准确性也越低，将严重影响聚类分析结果的可用性。另一方面，聚类中心的选取是随机的，没有明确标准，并不能保证选取结果是最优化的，也会影响聚类分析结果的稳定性，基于此文章提出一种改进的k-means算法。**

４.２　改进k-means聚类方法

**针对4.1节所述算法进行仿真实验，发现当隐私预算ε降低到一定值后，聚类的可用急剧下降，导致误差过大的重要原因之一在于初始中心点的选择，在随机选择中心点以后，添加了噪声的新中心点往往远远偏离了原中心点，因为对于原始中心点来讲，所添加的随机噪声会使新的中心点偏离很远，所以考虑将初始中心点的选取方法加以改进。**

**改进方法在挑选初始中心点时，采用将指数加噪机制与密度估计相结合的方法，挑选 K-这样可以保证初始中心挑选的合理性，同时也保障样本数据的隐私性。 算法的具体设计为：**

**输入：数据集 D，邻域半径 r，隐私预算，聚类的数目 K**

**输出：K 个簇**

**Step1 计算每一个样本数据的密度** ρ **；**

**Step2 将密度** ρ **小于阈值min 的样本数据删除，形成新的数据集 D'**

**step3 在数据集'D 中以正比于的概率输出一个初始中心点Ci ；**

**step4 将Ci 及其邻居点的密度**ρ**记为 0**

**step5 循环 2、3 步，直到选出第 K 个初始中心点**

**step6 给k个初始聚类添加噪声；**

**step7 针对数据集中的每个样本xi,计算它到k个聚类中心的距离并将其分到距离其最小的聚类中心所对应的类中，得到k个点的集合C={C1，C2…,Ck}；；**

**step8 针对每个集合Ci,计算各个点之和，和C集合中点数，并对sum与添加噪声，然后相除,并以此为新的中心点。**

**Step9 重复步骤2和步骤3直到聚类中心不再变化。**

**实验证明改进方法在维持噪声添加量和隐私保护级别 ε 不变的前提下大大提高了聚类的准确度。**

５实验分析

**文中涉及的算法均使用Java 语言进行仿真实现，实验环境具体为：Windows7操作系统 CPU2.6GHZ， 2.00G，MyEclipse6.0，算法所涉及的数据来自文献[6]，数据基本信息如表-１ 所示**

**表-1 数据基本信息表**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set** | **Dims** | **Type** | **Tuples** |
| **Unbalance** | **2** | **Real** | **6500** |
| **Gamma** | **11** | **Real** | **19020** |
| **ADL** | **3** | **Real** | **543** |

5.1 实验评价指标

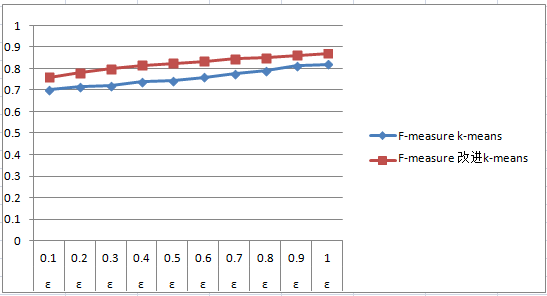
数据可用性可以通过两种方式进行评估:理论分析和实践验证。对于前者, (β, γ) -可用性[7]通常用于衡量差分隐私算法的可用性;对于后者，可用性度量包括相对误差、绝对误差、欧拉函数和F-measure[8]。聚类实验**采用评价指标Ｆ－ｍｅａｓｕｒｅ 来衡量。**与其他评价指标相比, F-measure的结果更具有针对性。

**下面介绍**F-measure**的计算方法，**假设n代表给定的数据集D的大小, i代表数据集的正确分类的类标签,和分别代表类i和簇Cj中的数据点的数量, nij代表类i和簇Cj的交集部分中的数据点的数量, 则**召回率和准确率的计算**公式定义如下:

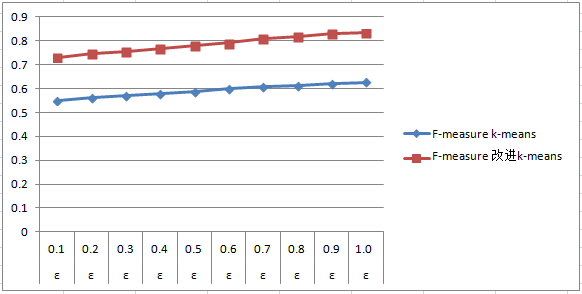
F-measure的计算公式定义如下:

5.2 实验结果

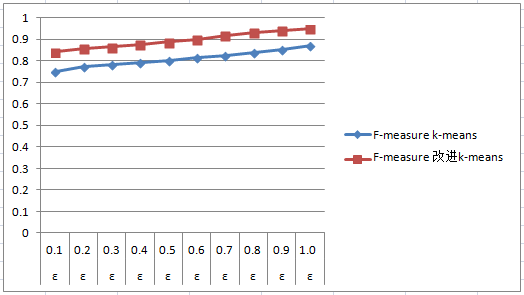
在2个数据集上分别运行差分隐私k-means算法和改进的差分隐私k-means算法，逐步将隐私预算ε的值从0调高到1.0。实验结果显示的是对应的每个ε值， 在2个数据集上分别调用2个算法聚类6次后得到的F-measure，具体结果如图1-图3：



**图-1数据集Unbalance上的运行结果**



**图-2数据集Gamma上的运行结果**



**图-3数据集ADL上的运行结果**

从聚类可用性分析，如图-1-图-所示, 在相同的ε值下, 改进的差分隐私k-means算法相比于为改进的差分隐私k-means算法具有更高的F-measure值, 更好地保证了聚类的可用性。随着差分隐私预算值的增加, F-measure值也随之增加, 这表明聚类结果随着隐私水平的降低而提高[8]。

**[1]** **方跃坚, 朱锦钟, 周文, 等.数据挖掘隐私保护算法研究综述[J].信息网络安全, 2017 (2) :6-11.**

**[2] Dework C, Lei J.Differential privacy and robust statistics[C]//** [**ACM symposium on theory of computing.[s.l.]:ACM, 2009:371-380.**](http://scholar.cnki.net/result.aspx?q=Differential%20privacy%20and%20robust%20statistics)

**[3]刘雅辉, 张铁赢, 靳小龙, 等.大数据时代个人隐私保护[J].计算机研究与发展, 2015, 52 (1) :1-19.**

**[4] BLUM A, DWORK C, MCSHERRY F, et al.Practical privacy:the SuLQ framework[C]∥Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.ACM,2005:128-138.**

**[5 DWORK C.Differential privacy[C]∥Proceedings of the 33rd International Conference on Automata, Languages and Programming-Volume Part II.Springer, Berlin, Heidelberg, 2006:1-19.**

**[6]FRNTI P.Clustering datasets[OL].http://cs.joensuu.fi/sipu/datasets.**

[**[7]NGUYEN H H, KIM J, KIM Y.Differential privacy in practice[J].Journal of Computing Science and Engineering, 2013, 7 (3) :177-186.**](http://scholar.cnki.net/result.aspx?q=Differential%20privacy%20in%20practice)

[**[8]JIANG H, YI S, LI J, et al.Ant clustering algorithm with Kharmonic means clustering[J].Expert Systems with Applications, 2010, 37 (12) :8679-868**](http://kns.cnki.net/kcms/detail/detail.aspx?dbcode=SJES&filename=SJES13011501640390&v=MTgxODhRSC9pclJkR2VycVFUTW53WmVadUh5bmdVcm5JSmw4V2JoYz1OaWZPZmJLN0h0RE5xbzlFWXU4UEQzVTVvQk1UNlQ0UA==&uid=WEEvREcwSlJHSldRa1FhcTdWa2JKYS9OblhIM0taRTJsa3Nlemc2Y2hsMD0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoHbKxJw!!)

**[9]van Rijsbergen C J.Information Retrieval (2nd edition) [M].London:Butterworths, 1979**