doi:10.11959/j.issn.1000-436x.2016038

MapReduce 框架下支持差分隐私保护的 *k*-means 聚类方法

李洪成[1]，吴晓平[1]，陈燕[2]
（1. 海军工程大学信息安全系，湖北 武汉 430033；2. 解放军 61062 部队，北京 100091）

摘 要：针对传统隐私保护方法无法应对任意背景知识下恶意分析的问题，提出了分布式环境下满足差分隐私的 *k*-means 算法。该算法利用 MapReduce 计算框架，由主任务控制 *k*-means 迭代执行；指派 Mapper 分任务独立并行计算各数据片中每条记录与聚类中心的距离并标记其属于的聚类；指派 Reducer 分任务计算同一聚类中的记录数量 *num* 和属性向量之和 *sum*，并利用 Laplace 机制产生的噪声扰动 *num* 和 *sum*，进而实现隐私保护。根据差分隐私的组合特性，从理论角度证明整个算法满足 *e*-差分隐私保护。实验结果证明了该方法在提高隐私性和时效性的情况下，保证了较好的可用性。

关键词：数据挖掘；*k*-均值聚类；MapReduce；差分隐私保护；Laplace 机制
中图分类号：TP301 **文献标识码：**A

k-means clustering method preserving differential privacy in MapReduce framework

LI Hong-cheng[1], WU Xiao-ping[1], CHEN Yan[2]
(1. Department of Information Security, Naval University of Engineering, Wuhan 430033, China;
2. No.61062 Troops of PLA, Beijing 100091, China)

Abstract: Aiming at the problem that traditional privacy preserving methods were unable to deal with malign analysis with arbitrary background knowledge, a *k*-means algorithm preserving differential privacy in distributed environment was proposed. This algorithm was under the computing framework of MapReduce. The host tasks were obligated to control the iterations of *k*-means. The Mapper tasks were appointed to compute the distances between all the records and clustering centers and to mark the records with the clusters which the records belong. The Reducer tasks were appointed to compute the numbers of records which belong to the same clusters and the sums of attributes vectors, and to disturb the numbers and the sums with noises made by Laplace mecha ism, in order to achieve differential privacy preserving. Based on the combinatorial features of differential privacy, theoretically prove that this algorithm is able to fulfill *e*-differentially private. The experimental results demonstrate that this method can remain available in the process of preserving privacy and improving efficiency.

Key words: data mining, *k*-means clustering, MapReduce, differential privacy preserving, Laplace mechanism

1 引言

数据挖掘作为信息获取的一种重要方法，可以从体量巨大、快速更新、类型多样、价值量大的大数据中挖掘出有用的信息。聚类分析是一种典型的非指导学习数据挖掘方法，主要思想是将数据分为若干类，使各聚类中的数据差别最小、聚类之间的数据差别最大，该方法在网络入侵异常检测、大规模选址和市场细分等领域有重要应用。

在大数据的背景下，聚类分析技术主要面临以下 2 个问题。1) 随着大数据时代的数据体量越来越巨大，单个计算机难以在可接受的时间内对数据进

收稿日期：2015-04-05；**修回日期：**2015-07-28
基金项目：国家自然科学基金资助项目（No.61100042）；总后军内科研基金资助项目（No.AWS14R013）
Foundation Items: The National Natural Science Foundation of China (No.61100042), The Military Scientific Research Project of the General Logistics Department (No.AWS14R013)

行有效的聚类分析。因此，如何利用并行分布式计算资源进行快速聚类分析[1]是亟待解决的关键问题。2) 数据聚类分析的结果在提供有价值信息的同时，可能会泄露数据集中单个记录的信息，对数据隐私安全造成威胁。在大数据时代，攻击者所拥有的背景知识越来越多，使攻击者窃取数据隐私更加便利[2]。因此，研究应对任意背景知识的隐私保护聚类分析技术成为隐私保护领域的研究焦点。

国内外学者做了许多卓有成效的研究工作。其中，文献[3]在云计算平台上采用高效的 MapReduce 并行计算模型实现了 *k*-means 聚类算法。该算法首先利用 MapReduce 的映射函数计算各条记录到聚类中心的距离，并标记其属于的聚类中心，然后利用规约函数计算出新的聚类中心，最后启用新的 MapReduce Job 来进行 *k*-means 算法的下一轮迭代，进而将 *k*-means 算法每轮迭代中时间复杂度最高的步骤交由分布式计算资源处理，有效提高了 *k*-means 算法的运行效率。在提高聚类分析时效性的同时，云平台的开放性使攻击者拥有大量的攻击背景知识[4]，攻击者可以通过关联背景知识和聚类结果来窃取数据隐私[5,6]。然而，传统的隐私保护方法只能应对特定背景知识下的攻击，针对此问题，文献[7]利用差分隐私保护的严格定义，提出了可以应对任意背景知识下攻击的 *k*-means 聚类方法 DP *k*-means（differential private *k*-means），该方法通过对 *k*-means 算法每轮迭代中各聚类内记录之和 *sum* 和记录数 *num* 等中间变量加入适量随机噪声来实现隐私保护。此外，文献[8,9]针对 DP *k*-means 中初始聚类中心受随机噪声影响较大的问题，对 DP *k*-means 算法的初始中心点选择方法进行了改进，有效提高了聚类分析的可用性。以上文献并没有研究将 DP *k*-means 部署于分布式环境的具体方法，也就没有解决分布式环境下应对任意背景知识的数据隐私保护问题。

基于此，本文提出了一种 MapReduce 框架[10]下支持差分隐私保护的 *k*-means 聚类方法，利用 MapReduce 框架提供的分布式计算功能来提高聚类分析的效率，并通过随机噪声添加使聚类的输出结果满足差分隐私。

2　差分隐私保护

差分隐私保护是一种基于部分信息隐藏的隐私保护技术，该技术通过随机响应或随机噪声添加来实现信息干扰，同时使干扰后的输出信息在一定程度上保持原有的统计特性，进而使数据挖掘结果的可用性保持在可接受的范围内。

差分隐私技术给出了一个严格且可证明的隐私保护定义，保证在数据集中改变任一条记录时，查询结果的变化量极小，攻击者在已知除目标记录外所有其他记录信息的条件下，仍然无法分析出这条记录的任何信息，因此该方法可以应对任意背景知识下的恶意分析[11]。差分隐私保护的基本原理如下。

用户从数据集 D 中提取信息的操作被定义为查询 F，算法 A 对查询 F 的输出进行随机化处理，使之满足差分隐私保护的条件[12]。

定理 1　假设数据集 D 和 D' 完全相同或只相差一条记录，$Range(A)$为一个随机算法 A 输出的值域，$Pr[X]$为事件 X 发生的可能性，如果对于任意 $S\in Range(A)$，有

$$Pr[A(D)\subseteq R_A]\leq e^{e}\,Pr[A(D')\subseteq R_A] \tag{1}$$

则随机算法 A 提供 e-差分隐私保护，其中，参数 e 称为隐私保护预算。

全局敏感度是查询函数的一条重要固有属性，反映单个记录变化对查询函数输出的影响。全局敏感度的定义如下。

定义 1　查询 F 的全局敏感度为

$$\Delta F=\max_{D,D'}\left\|F(D)-F(D')\right\|_1 \tag{2}$$

其中，$\|\ \|_1$ 表示向量各元素的绝对值之和，数据集 D 和 D' 完全相同或只相差一条记录，d 是数据集中记录的属性维数。

差分隐私保护的实现机制主要有 Laplace 机制与指数机制 2 种，分别利用随机噪声添加和随机响应的方式实现隐私保护。其中，Laplace 机制适用于对数值型结果的保护[13]，是聚类分析中最常用的差分隐私保护机制，其原理如下。

定理 2　对于查询 F，数据集 D，设查询输出为 $F(D)$，F 的全局敏感度为 $?F$，如果噪声 Y 服从尺度为 $\frac{\Delta F}{e}$ 的拉普拉斯分布，则算法 $A(D)=F(D)+Y$ 满足 e-差分隐私[13]。

在定理 2 中，随机噪声 $Lap\left(\frac{\Delta F}{e}\right)$ 的概率密度函数为

$$p(x)=\frac{1}{2\left(\frac{\Delta F}{e}\right)}\exp\left[-\frac{|x|}{\left(\frac{\Delta F}{e}\right)}\right] \tag{3}$$

此外，差分隐私保护具有序列组合性和并行组合性 2 种组合特性，这些特性在证明算法是否满足差分隐私以及在隐私预算分配过程中起着重要作用[14]。

性质 1 设有 m 个随机算法 $A_1,\cdots,A_m$，算法 A_i $(1\leq i\leq m)$提供 e_i-差分隐私保护，则对于同一数据集 D，$\{A_1,\cdots,A_m\}$在 D 上的序列组合算法提供 e-差分隐私保护，其中，$e=\sum_{i=1}^{m}e_i$。

性质 2 设有随机算法 A 和数据集 D，将 D 分为不相交的子集 $D_1,\cdots,D_n$，若算法 A 提供 e-差分隐私保护，则 A 在$\{D_1,\cdots,D_n\}$上的组合运算所构成的算法提供 e-差分隐私保护。

3 MapReduce 框架下的 DP *k*-means 算法

本文算法的功能是在 MapReduce 分布式环境下，保证在数据集中改变任一记录时，每个聚类的质心以及记录数量所发生的变化不泄露隐私信息，即恶意分析者无法利用其拥有的与原数据集相似的数据集，通过挖掘得到原数据集中单个记录的隐私信息。算法应对的攻击模型如图 1 所示。

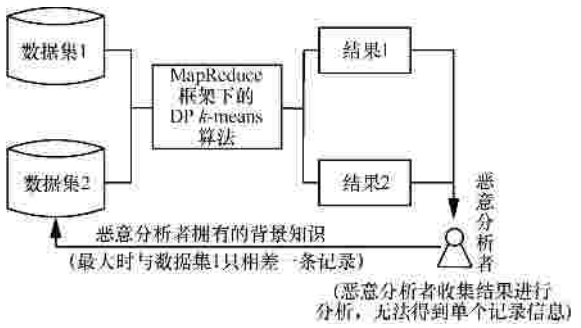


图 1 本文算法应对的攻击模型

算法的基本思路是利用分布式计算节点上的 Map 分任务判断出各记录属于的聚类类别，利用 Reduce 分任务计算出聚类中的记录数量和对应属性之和，并加入适量 Laplace 噪声，使聚类分析的结果满足 e-差分隐私。

传统的差分隐私保护 k-means 算法[7]存在聚类准确性较低的问题，其主要原因是随机选择的初始中心点导致算法收敛速度较慢，而算法迭代次数的增多导致了每轮添加的噪声增多，而且添加了噪声的初始中心点往往与原中心点偏离较远[9]，这些问题均导致了算法的聚类准确性较低，所以本算法在对差分隐私保护 k-means 算法进行 MapReduce 并行化设计时，采用改进的初始中心点选择和加噪方法。

3.1 算法设计

设数据集中的记录总数为 N，各条记录记为 a_i $(1\leq i\leq N)$，各记录的维数为 d；将这些记录分为 M 个数据片，各数据片记为 D_j $(1\leq j\leq M)$；算法要求的聚类数目为 K，各聚类中心记为 u_k $(1\leq k\leq K)$。算法的步骤如下。

Step1 主任务 Driver 首先将各记录归一化到 $[0,\ 1]^d$ 空间中。将 N 条记录 $a_1,\cdots,a_N$ 平均分成 K 个子集 $C_1,\cdots,C_K$，集合 C_k 中的记录数$|C_k|\leq ceil\left(\frac{N}{K}\right)$，$ceil()$为向上取整函数[8]。计算 C_k 中记录的数量 num_k^0 和 C_k 中各记录的属性向量之和 sum_k^0，分别对 num_k^0 和 sum_k^0 加入随机噪声得到 $num_{k'}^0$ 和 $sum_{k'}^0$，计算 $u^0{}_{k}{}'=\left(\frac{sum_{k'}^0}{num_{k'}^0}\right)$，$u_{k'}^0$ 即为初始聚类中心点。

Step2 主任务将所有数据记录平均分为 M 个数据片，并指派 M 个分任务执行 Map 操作，指派 K 个分任务执行 Reduce 操作。

Step3 Mapper 分任务接收包含 $\frac{N}{M}$ 个记录的数据片，运行 Map 函数：计算每个记录到聚类中心的距离，并选择距离最小的聚类中心。每条记录得到<*key*，*value*>对的格式为<该记录属于的聚类中心标识，该记录的属性向量>。

Step4 Reducer 分任务接收所有同属于一个聚类中心的<*key*，*value*>对，运行 Reduce 函数：计算该聚类中记录的数量 *num* 和聚类内各记录的属性向量之和 *sum*，对 *num* 和 *sum* 加入随机噪声，然后计算加噪后的聚类中心点 u'。

Step5 主任务接收各个 Reduce 节点的输出结果 u'，计算本轮和上一轮中 K 个聚类中心点的距离。若中心点属性向量差的距离范数小于阈值，则算法终止，输出各聚类中心和聚类内记录的数量；否则，重复 Step3~Step5。

MapReduce 框架下的 DP *k*-means 算法流程如图 2 所示。

3.2 隐私性分析

由图 2 可以看出，MapReduce 框架下 DP *k*-means 算法的隐私性通过对每个 Reduce 操作中的 num_k 和 sum_k 加入 Laplace 噪声来实现。由于 *k*-means 算法的每轮迭代相当于随机算法的序列组合，所以

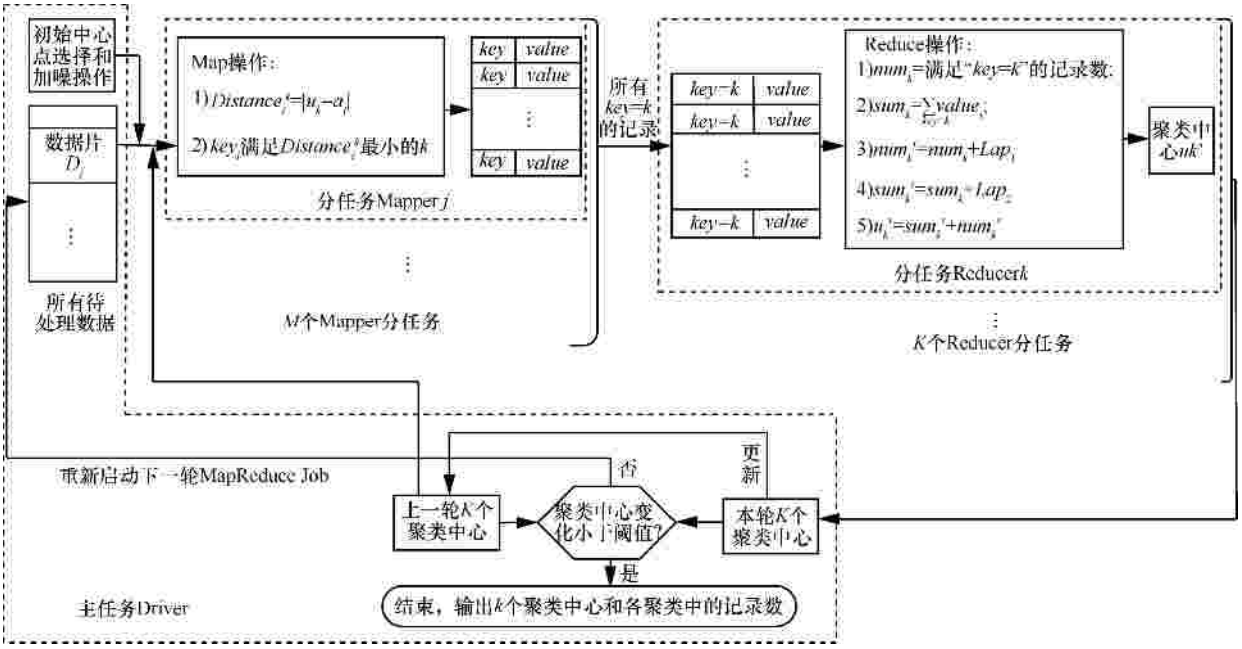


图 2　MapReduce 框架下的 DP *k*-means 算法

根据性质 1 可知，整个算法的隐私保护预算为

$$e=\sum_{t=1}^{T}e_t \tag{4}$$

其中，T 为迭代的总次数，e_t 为第 t 次迭代的隐私保护预算。在预算分配方面，本文采用文献[7]的策略，每次迭代消耗剩余隐私预算的一半，即第 t 次迭代的隐私预算为 $e_t=\frac{e}{2^t}$。

在每轮迭代中，由于 K 个 Reducer 分节点相互独立地执行操作，各轮迭代的结果相当于 Reduce 操作的并行组合[15]，所以根据性质 2 可知，要想第 t 轮迭代满足 e_t-差分隐私，需要使分布式环境下每个 Reducer 分任务的操作满足 e_t-差分隐私。

由定义 1 可知，*num* 的全局敏感度 $?f_{num}=1$，在 d 维空间$[0, 1]^d$ 的点集中添加或删除一个点，各个属性和的最大变化量为 1，由于点的维数为 d，则 *sum* 的全局敏感度 $?f_{sum}=d$。所以根据性质 1 可知，整个查询序列的全局敏感度 $?f=d+1$。

因此，由定理 2 可知，在初始中心点计算过程中的 num^0_k 和 sum^0_k 上分别加入随机噪声 $Lap\ (d+1)\frac{2}{e}$，并在算法第 t 轮迭代的 num_k 和 sum_k 上分别加入随机噪声 $Lap(d+1)\frac{2^{t+1}}{e}$，可以保证 MapReduce 框架下 DP *k*-means 算法满足 e-差分隐私保护。相比于传统的差分隐私保护 *k*-means 算法，改进的初始中心点计算过程可以在相同的隐私保护预算下，减少算法的迭代次数，进而减少随机噪声添加量。

4　算法效率及可用性实验

由于本文算法的主要功能是利用 MapReduce 分布式计算框架提高聚类效率，并利用 Laplace 机制保护数据隐私，而算法的隐私性已经在上文中得到证明，因此本文的实验部分只考虑算法的运行效率，以及隐私保护聚类算法的输出结果可用性。

实验中云计算平台由 1 台主节点的计算机和 3 台分节点的计算机组成，每台计算机配置如下：操作系统为 Linux，CPU 为 3.30 GHz，内存为 2.99 GB。在集群上部署 Hadoop0.20.2。聚类算法利用 Java 软件进行开发。

实验所选择的数据集为 UCI Knowledge Discovery Archive database 中的“Blood”数据集（记录数为 748，属性数为 5，数据类型为实值型）和“gamma”数据集（记录数为 19 020，属性数为 10，数据类型为实值型），这 2 个数据集中均给出了各记录的分类信息，因此可以用来检验聚类算法的性能。根据“Blood”和“gamma”数据集的标准分类结果，实验设定聚类中心数为 2，相邻 2 轮中心点属性向量差的距离范数小于 1 时迭代终止。

4.1　算法运行效率实验

为反映 MapReduce 框架下算法运行效率受分布式计算节点数量的影响，本实验启动 1 个和 3 个

分节点分别进行运算，并考察不同数据量情况下加速比的变化规律。首先，在“gamma”数据集上截取不同数量的记录作为待处理文件，分别上传至 Hadoop 的 HDFS 文件系统中。然后，将本文算法部署于 MapReduce 中，记录 5 次运行时间的平均值，得到 1 个和 3 个子节点分别参与运算的情况下算法的运行时间（如图 3 所示）。

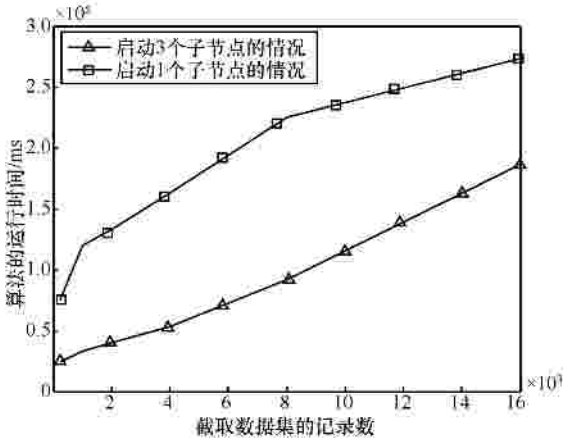


图 3 算法的运行时间

由图 3 可以看出，系统启动 3 个子节点时算法的运行时间较启动 1 个子节点时显著减少，说明分布式集群可以有效提高本文聚类算法的运行效率。然而，随着数据集记录数量的增多，运行效率的提高比例逐渐降低。主要原因是记录数量的增涨导致算法迭代次数增加，而每轮迭代中都有主节点与子节点之间的数据传递操作，以及主节点进行的聚类中心距离比较操作，这些操作没有利用分布式集群的资源，因此操作用时随着迭代次数增加而增加，而不随子节点数目增加而减少。

另外，为反映本文算法引入隐私保护后对算法效率的影响，将本文算法与分布式计算框架下的 *k*-means 算法进行比较。由于文献[3]算法将 *k*-means 算法部署于 MapReduce 环境中，而没有引入隐私保护，所以选择文献[3]算法作为比较对象。首先，在包含 1 个主节点和 3 个分节点的分布式计算平台上，分别利用本文算法和文献[3]算法处理“gamma”数据集上截取的不同文件，记录 5 次运行时间的平均值，得到本文算法运行时间减去文献[3]算法运行时间的差值（如图 4 所示）。

通过比较图 3 和图 4 中数值的数量级，可以看出 2 种算法运行时间的差值比算法运行时间要小约 4 个数量级。因此，相比于文献[3]算法，本文算法在提高隐私性的情况下，并没有导致运行效率的明显降低。此外，由图 4 可以看出，2 种算法运行时间的差值随着数据集记录数量的增涨而增大。主要原因是随着记录数量的增长，算法的迭代次数逐渐增加，而本文算法的每轮迭代都要进行 Laplace 随机噪声的产生和添加操作，每轮迭代中的这些操作直接导致了 2 种算法运行时间的差别。

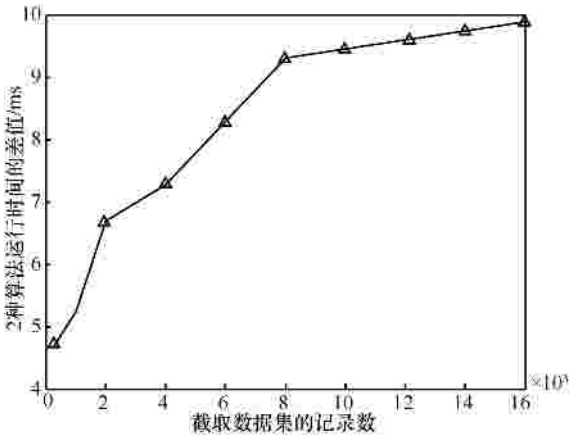


图 4 本文算法与文献[3]算法运行时间的差值

4.2 聚类结果可用性实验

为衡量本文提出的 MapReduce 框架下差分隐私聚类算法的可用性，本实验选择以下参考标准与本文算法聚类结果进行比较：1) 由于“Blood”和“gamma”数据集中的标准分类情况是在真实情况下调查得到的，因此选择“Blood”和“gamma”数据集的标准分类结果作为比较对象之一；2) 为了分析本文算法引入隐私保护后对聚类结果可用性的影响，本实验选择文献[3]算法的聚类结果作为可用性比较的另一个参考对象。

衡量数据挖掘结果可用性的指标主要有准确率（precision）和召回率（recall）等，而 F-measure 可以对准确率和召回率进行综合，因此本实验利用 F-measure 评价指标来衡量聚类可用性。F-measure 越大说明 2 个聚类结果的相似程度越强，即本文算法所添加的噪声对聚类可用性影响越小。

F-measure 的计算过程如下：用 *CLUSTER* 表示作为参考标准的聚类结果，*CLUSTER'* 表示本文聚类算法的聚类结果，聚类数为 K，U_i 为 *CLUSTER* 中的第 i 个聚类集合（$1 \le i \le K$），V_i 为 *CLUSTER'* 中的第 i 个聚类集合（设 2 次聚类的标记统一），$cover_i$ 为 U_i 和 V_i 重合的记录数目，$|U_i|$和$|V_i|$分别为 U_i 和 V_i 中的记录数目，记第 i 个聚类的准确率为 P_i，召回率为 R_i，则

$$R_i = \frac{cover_i}{|U_i|} \tag{5}$$

$$P_i = \frac{cover_i}{|V_i|} \tag{6}$$

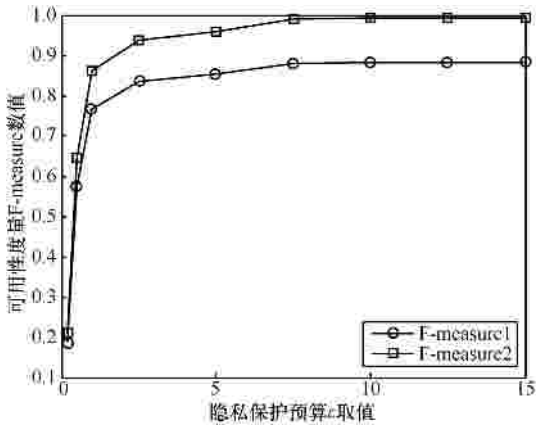
然后，计算 P_i 和 R_i 的加权调和平均，记为 F_i，则

$$F_i = \frac{2R_iP_i}{R_i + P_i} \tag{7}$$

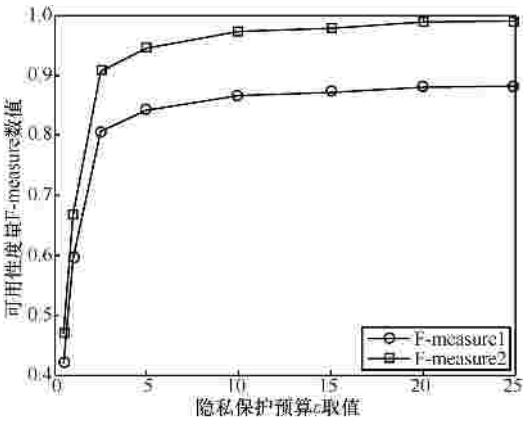
最后，对各聚类的 F_i 进行加权平均。设 N_{TOTAL} 为数据集记录总数，算得聚类结果的可用性度量

$$F\text{-}measure = \sum_{U_i \in CLUSTER} \frac{|U_i|}{N_{\mathrm{TOTAL}}} F_i \tag{8}$$

将本文算法的聚类结果与数据集中标准分类结果的相似程度记为 F-measure1，并将本文算法的聚类结果与文献[3]算法聚类结果的相似程度记为 F-measure2。由于算法每次运行时噪声的添加量服从 Laplace 随机分布，所以聚类结果具有一定的随机性，因此实验中的结果取 10 次运算的平均值。对于“Blood”和“gamma”数据集，当隐私保护预算 e 变化时，F-measure1 和 F-measure2 的变化情况如图 5 所示。



(a)“Blood”数据集聚类结果的 F-measure 变化情况



(b)“gamma”数据集聚类结果的 F-measure 变化情况

图 5　各数据集的 F-measure1 和 F-measure2 随 e 变化情况

由图 5 可以看出，当隐私保护预算 e 大于 3 时，本文算法的聚类结果可用性达到较高水平，因此本文算法可以在实现隐私性的情况下，保证聚类结果具有较好的可用性。另外，图 4 中当隐私保护预算 e 较小时，聚类结果的可用性随着隐私保护预算的增加而显著增加，当 e 增大到一定值时，聚类结果可用性的增加速度趋于平缓，而考虑到隐私保护预算 e 与添加噪声的尺度参数成反比，因此用本文算法处理“Blood”和“gamma”数据集时，隐私保护预算 e 取 3~4 有利于实现算法隐私性和可用性的平衡。

5　结束语

本文利用 MapReduce 计算框架实现了并行分布式 *k*-means 聚类，并同时利用 Laplace 机制实现了算法的差分隐私保护，同时提高了 *k*-means 算法的时效性和隐私性。下一步需要进行的研究工作主要有以下 2 个方面：1) 针对算法迭代次数对运行效率影响较大的问题，设计算法减少 MapReduce Job 的运行次数，并将 *k*-means 中的迭代执行判断工作交由子节点执行；2) 为缓解算法隐私性和可用性之间的矛盾，研究如何在保证隐私保护水平的前提下，通过综合利用指数机制和 Laplace 机制，减少对每轮迭代实施的扰动。

参考文献：

[1] FLAVIO C, NILESH D, RAVI K. Correlation clustering in MapReduce[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014). New York, USA, c2014: 641-650.

[2] 孟小峰，张啸剑. 大数据隐私管理[J]. 计算机研究与发展，2015, 52(2):265-281.
MENG X F, ZHANG X J. Big data privacy management[J]. Journa f Computer Research and Development, 2015, 52(2):265-281.

[3] 江小平，李成华，向文，等. *k*-means 聚类算法的 MapReduce 并行化实现[J]. 华中科技大学学报（自然科学版），2011, 39(S1): 120-124.
JIANG X P, LI C H, XIANG W, et al. Parallel implementing *k*-means clustering algorithm using MapReduce programming mode[J]. nal of Huazhong Univ of Sci & Tech(Natural Science Edition), 2011, 39(S1):120-124.

[4] ROY I, SETTY S T V, KILZER A, et al. Airavat: security and privacy for MapReduce[C]//The 7th USENIX Symposium on Networked Systems Design and Implementation. San Jose, USA, c2010:297-312.

[5] 肖人毅. 云计算中数据隐私保护研究进展[J]. 通信学报，2014, 35(12):168-177.
XIAO R Y. Survey of privacy preserving data queries in cloud computing[J]. Journal on Communications, 2014, 35(12):168-177.

[6] SHI E, CHAN T H, RIEFFEL E G, *et al*. Privacy-preserving aggregation of time-series data[C]//The Network and Distributed System Se-

curity Symposium. San Diego, USA, c2011.

[7] DWORK C. A Firm Foundation for Private Data Analysis[J]. Communications of the ACM, 2011, 54(1):86-95.

[8] 漕?，??香 ，?岷? ，等. 差分隐私 DPE *k*-means 数据聚合下的多维数据可视化[J]. 小型微型计算机系统, 2013, 34(7):1637-1640.
LI Y, HAO Z F, XIAO Y S, et al. Multidimensional data visualization using aggregation method of differential privacy equip partition k-means[J]. Journal of Chinese Computer Systems, 2013, 34(7): 1637-1640.

[9] 漕?，??香 ，?? ，等. 差分隐私保护 *k*-means 聚类方法研究[J]. 计算机科学, 2013, 40(3):287-290.
LI Y, HAO Z F, WEN W, et al. Research on differential privacy preserving *k*-means clustering[J]. Computer Science, 2013, 40(3): 287-290.

[10] 何清, 庄福振, 曾立, 等. PDMiner: 基于云计算的并行分布式数据挖掘工具平台[J]. 中国科学: 信息科学, 2014, 44(7):871-885.
HE Q, ZHUANG F Z, ZENG L, et al. PDMiner: a cloud computing based parallel and distributed data mining toolkit platform[J]. Chinese Science: Information Science, 2014, 44(7):871-885.

[11] MCGREGOR A, MIRONOV I, PITASSI T, et al. The limits of two-party differential privacy[C]//The 51st IEEE Annual Symposium on Foundations of Computer Science. Las Vegas, USA, c2010:81-90.

[12] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1):101-122.
XIONG P, ZHU T Q, WANG X F. A survey on differential privacy and applications[J]. Chinese Journal of Computers, 2014, 37(1):101-122.

[13] 丁丽萍, 卢国庆. 面向频繁模式挖掘的差分隐私保护研究综述[J]. 通信学报, 2014, 35(10):200-209.
DING L P, LU G Q. Survey of differential privacy in frequent pattern mining[J]. Journal on Communications, 2014, 35(10):200-209.

[14] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014, 37(4):927-949.
ZHANG X J, MENG X F. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers, 2014, 37(4):927-949.

[15] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[J]. Communication of the ACM, 2010, 53(9):89-97.

作者简介：

李洪成（1991-），男，河南商丘人，海军工程大学博士生，主要研究方向为信息安全、数据挖掘。

吴晓平（1961-），男，山西新绛人，海军工程大学教授、博士生导师，主要研究方向为信息安全、密码学。

陈燕（1975-），女，河北石家庄人，解放军 61062 部队高级工程师，主要研究方向为网络应用、信息系统。