1. 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(01):101-122.
   1. 背景

差分隐私是Dwork在2006年针对统计数据库的隐私泄露问题提出的一种新的隐私定义。在此定义下，对数据集的计算处理结果对于具体某个记录的变化是不敏感的，单个记录在数据集中或者不在数据集中，对计算结果的影响微乎其微。所以，一个记录因其加入到数据集中所产生的隐私泄露风险被控制在极小的、可接受的范围内，攻击者无法通过观察计算结果而获取准确的个体信息。比如当数据集Ｄ 中包含个体Alice时，设对D进行任意查询操作f(例如计数、求和、平均值、中位数或其它范围查询等)所得到的结果为f(D)，如果将Alice的信息从Ｄ 中删除后进行查询得到的结果仍然为f(D)，则可以认为，Alice的信息并没有因为被包含在数据集Ｄ中而产生额外的风险。差分隐私保护就是要保证任一个体在数据集中或者不在数据集中时，对最终发布的查询结果几乎没有影响。具体地说，设有两个几乎完全相同的数据集（两者的区别仅在于一个记录不同），分别对这两个数据集进行查询访问，同一查询在两个数据集上产生同一结果的概率的比值接近于１。

差分隐私保护能解决传统隐私保护模型的两个缺陷。首先，差分隐私保护模型假设攻击者能够获得除目标记录外所有其它记录的信息，这些信息的总和可以理解为攻击者所能掌握的最大背景知识。在这一最大背景知识假设下，差分隐私保护无需考虑攻击者所拥有的任何可能的背景知识，因为这些背景知识不可能提供比最大背景知识更丰富的信息。其次，它建立在坚实的数学基础之上，对隐私保护进行了严格的定义并提供了量化评估方法，使得不同参数处理下的数据集所提供的隐私保护水平具有可比较性。

差分隐私保护模型最初被应用在统计数据库安全领域，旨在发布统计信息时保护数据库中个体的隐私信息，之后被广泛应用于隐私保护数据发布(Privacy Preserving Data Release, PPDR)与隐私保护数据挖掘(Privacy Preserving Data Mining, PPDM)等领域。

* 1. 差分隐私理论基础
     1. 基本定义

**定义1：**差分隐私. 对于一个随机算法M，*Pm*为算法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*Dʹ*以及*Pm*的任何子集*Sm*，若算法M满足，则称算法M提供*ɛ-*差分隐私保护，其中ɛ称为隐私保护预算。

* + 1. 相关概念

1. 隐私保护预算

隐私保护预算ε用来控制算法Ｍ在两个邻近数据集上获得相同输出的概率比值，它事实上体现了Ｍ所能够提供的隐私保护水平。在实际应用中，ε通常取很小的值，例如0.01，0.1，或者ln2，ln3等。ε越小，表示隐私保护水平越高．当ε等于０时，保护水平达到最高，此时对于任意邻近数据集，算法都将输出两个概率分布完全相同的结果，这些结果也不能反映任何关于数据集的有用的信息。因此，ε的取值要结合具体需求来达到输出结果的安全性与可用性的平衡。

1. 敏感度

噪声过小无法保护隐私，过量影响结果可用性。敏感度是决定噪声量大小的关键参数，它指删除数据集中任一记录对查询结果造成的最大改变。差分隐私保护方法中定义两种敏感度：全局敏感度和局部敏感度。

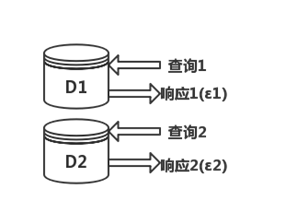
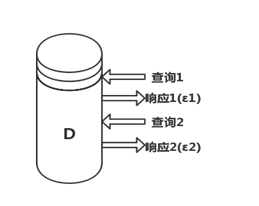
**定义2：**全局敏感度.设有函数*f*:*D*→*Rd*，D为输入数据集，输出为*d*维向量。对于相邻数据集*D*和*Dʹ*，有称为函数*f*的全局敏感度，其中是和之间的1-阶范数距离。

**定义3：**局部敏感度. 设有函数*f*:*D*→*Rd*，D为输入数据集，输出为*d*维向量。对于相邻数据集*D*和*Dʹ*，有称为函数*f*在D上的局部敏感度。

函数全局敏感度由函数本身决定，不同的函数有不同的全局敏感度，例如计数函数，全局敏感度为1。求均值和求中位数等函数的敏感度一般较大。局部敏感度由函数f和数据集D中的具体数据共同决定。

* + 1. 组合性质

一个复杂的隐私保护问题通常需要多次应用差分隐私保护算法才能得以解决。在这种情况下，为了保证整个过程的隐私保护水平控制在给定的预算ε之内，需要合理地将全部预算分配到整个算法的各个步骤中，这时可以利用隐私保护算法的两个组合性质，如图所示。



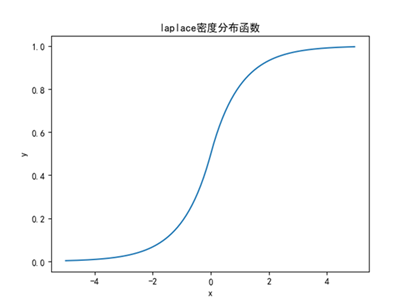
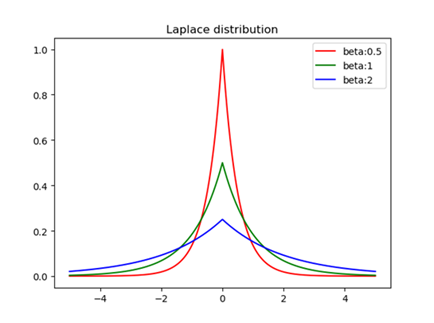
(a)序列组合性 (b)并行组合性

**性质1：**序列组合性.设有算法*M1*,*M2*,…*Mn*，其隐私保护预算分别为ɛ1,ɛ2,…,ɛn，那么对于同一数据集Ｄ，由这些算法构成的组合算法*Ｍ*(*Ｍ1*(*D*)*,Ｍ2*(*D*)*,…，Ｍn*(*D*))提供-差分隐私保护。

**性质2：**并行组合性. 设有算法*M1*,*M2*,…*Mn*，其隐私保护预算分别为ɛ1,ɛ2,…,ɛn，那么对于不相交的数据集*D1*,*D2*,…*Dn*，由这些算法构成的组合算法*Ｍ*(*Ｍ1*(*D*)*,Ｍ2*(*D*)*,…,Ｍn*(*D*))提供max*ɛi*-差分隐私保护。

* + 1. Laplace机制

Laplace机制适用于对数值数据保护，通过向确切的查询结果中加入服从Laplace分布的随机噪声来实现差分隐私。记位置参数为0、尺度参数为*b*的Laplace分布为*Lap*(*b*)，那么其概率密度为，期望是0，方差为2*b2*，*b*表示图像宽窄的尺度。



**定义4：**Laplace机制.设有函数f:D→Rd，D为给定数据集，其敏感度为∆*f*，那么随机算法*M*(*D*)*=f*(*D*)+*Y*提供ɛ-差分隐私保护，其中*Y~*为随机噪声，可以看出ɛ越小，引入的噪声越大。

1. 康海燕,马跃雷.差分隐私保护在数据挖掘中应用综述[J].山东大学学报(理学版),2017,52(03):16-23+31.
   1. 基于差分隐私的模式挖掘
2. DP-top k频繁模式挖掘

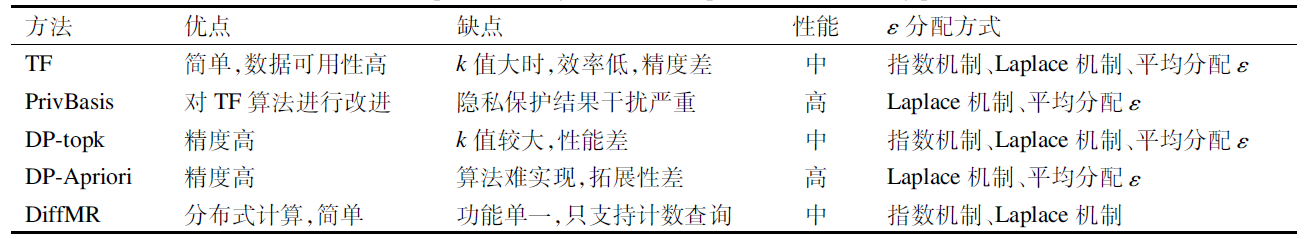
结合了Laplace和指数机制，满足了差分隐私保护的要求，对支持度计数进行后置处理。首先使用FP-Growth方法挖掘频繁模式集合，采用指数机制从中选出top *k*个频繁模式，进一步对*k*个模式的支持度添加Laplace噪音，对数据进行一致性处理，提升了可用性。

1. DiffMR

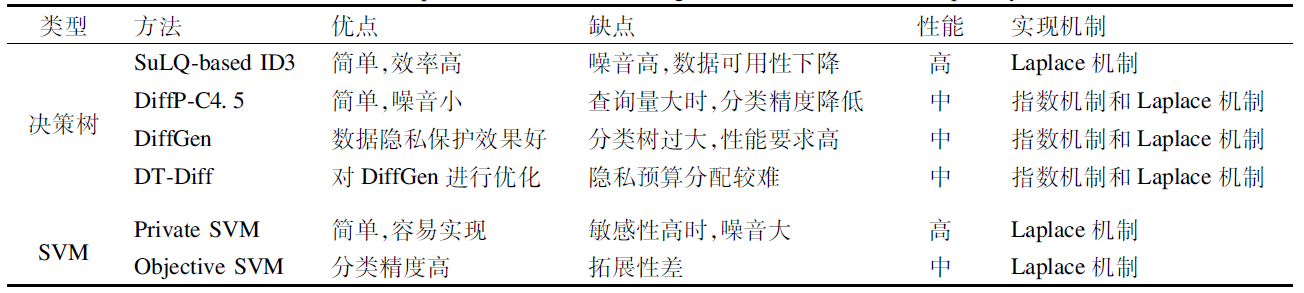
DiffMR(Differentially private Top-*k* query over MapReduce)用来解决MapReduce大数据分析查询中的隐私泄露问题。使用指数机制和打分函数选出top *k*个记录，当得到较精确的结果之后，对每条数据添加Laplace噪音采用后置处理精确查询结果。该算法仅支持简单的计数查询，ｋ值较大时，为了得到精确的结果，性能和效率都会降低。

1. DP-Aprior

首先变换数据库限制长度（智能权重截取算法），算法的主要思想是利用候选频繁项集引导事务截取操作，使在同一子事务重点的项集更容易产生频繁项集；然后在变换数据中计算每个候选集的噪音支持度，估算原始数据库中的噪音支持度（支持度估算算法）。进行阈值对比，生成合适的候选项集，应用ε-差分隐私FIM算法。



* 1. 基于差分隐私的分类



* 1. 基于差分隐私的聚类

Dwork针对*k-*means算法进行了改进，当迭代次数为Ｎ时，使用噪声分布*Lap*((d+1)N/ε)。当迭代次数不确定时，每次迭代使用上次剩余迭代隐私预算一半。当ε的值较小时，添加噪声后的中心点与随机选择的中心点有偏差，导致聚类分析准确度会有较大幅度下降。IDP *k-*means聚类算法，通过对加扰的子集中心点进行处理，提高了聚类准确度并通过实验进行了验证。在处理大数据集时，差分隐私聚类算法还没有提供一个高效的途径选择ｋ值。

聚类技术被归纳为一种非监督式学习，主要应用于机器学习、数据挖掘、模式识别、图像分析和生物信息等。本文分析了聚类隐私保护的若干技术，如何选取合理的Ｋ值是隐私保护和聚类准确度的关键，因此选取Ｋ值是未来的研究方向。