http://www.52ml.net/15084.html刚刚的想法！这篇文章是说GBDT的，是否可以像那个把差分隐私应用在随机森林中一样，把差分隐私应用在GBDT中。。回来想一下。

这周主要读了Differential Private Random Forest这篇文章，文章选择了在随机森林分类器中应用差分隐私。文章的第一部分是在说明保护隐私的重要性，为什么选择随机森林（比其他决策树算法有更好的精度；使用随机森林主要优点是不需要剪枝，因此计算效率高），以及文章的组织结构。第二部分讨论了差分隐私的一些相关工作，这里主要介绍了Arik Friedman 和Assaf Schuster提出的决策树中应用差分隐私的方法，认为他们主要的限制是打分函数的选择，这个影响了最后分类的准确性。并且认为Differential Private ID3比SuLQ-based ID3好。在本文中将采用Differential Private ID3算法。理论背景部分，介绍了差分隐私概念，敏感度，laplase机制，指数机制，以及随机森林的相关概念。主体部分在接下来会详细看。目前看到14年的几篇关于分类和差分隐私相关的文章，都在关注随机森林，所以可能最近会考虑是不是从这边入手。

Differential Private Random Forest

在我们的工作中，我们要研究是否有可能将差分隐私应用到决策树算法中，并且有好的准确性和高的效率。在我们的工作中，我们考虑使用随机森林[ 3 ]提供差分隐私问题，具有比其他决策树算法更好的精度。随机森林在大的数据集上运行高效，没有删除，并且估计了哪些变量是分类中重要的。此外，使用提供差分隐私的随机森林主要优点是不需要剪枝，这在大多数的基于决策树算法中需要，因此它的计算效率更高。本文的研究，不仅工作在差分隐私框架的随机森林，而且对不同的打分函数，如信息增益，最大算子，基尼指数，研究在结果合适时的噪声对准确性和敏感性的影响。每当对手分析差分私人随机森林的结果，他看不到任何私人信息。这个过程如图1所示。

本文的其余部分组织如下。第二部分，讨论了几个相关的工作在差分隐私。第三，介绍理论背景差分隐私，指数机制和随机森林。第四，说明提出的微分私人的随机森林算法的工作。第五部分给出了实验结果和第六节讨论这些结果。最后，第七节总结全文。



第一部分是在说明保护隐私的重要性；以及为什么选择随机森林（比其他决策树算法有更好的精度；不需要剪枝，计算效率高）；文章的组织结构。

相关工作这一部分

Arik Friedman and Assaf Schuster [ 5 ]，提出了利用决策树归纳为例差分隐私数据挖掘。他们认为，正式的隐私保障纳入一个系统介绍数据挖掘需要采取一种不同的方法的数据挖掘算法。他们工作的主要缺陷是敏感质量功能如信息增益[ 4 ]和[ 5 ]最大算子对噪声，从而影响最后的分类精度。他们吃不同的质量等功能的信息增益，最大算子和基尼指数为指数机制，发现了一些有趣的实验结果。他们的研究结果表明，不同质量的使用功能，影响最终结果的准确性。理由是，这些功能由于噪声的敏感性；信息增益比max算子对噪声比较敏感

Differential Private ID3比SuLQ-based ID3好。

Differential Private ID3在本文中将采用该算法。

理论背景部分，介绍了差分隐私概念，敏感度，laplase机制，指数机制；以及随机森林

**IV. DIFFERENTIAL PRIVATE RANDOM FOREST**

这一部分比较重要。

把差分隐私的概念应用到经典随机森林算法中。为了使随机森林隐私化，我们使用了Differential private ID3在构建随机森林的每一个迭代过程中。使用指数机制从m个属性中选择分裂属性。

input 数据集D，有属性A（d个）和一个分类属性C

先前的工作文献15中提到。。以前只有categorical datasets，现在扩展到continuous datasets，于是需要预处理连续数据集。

**A预处理**

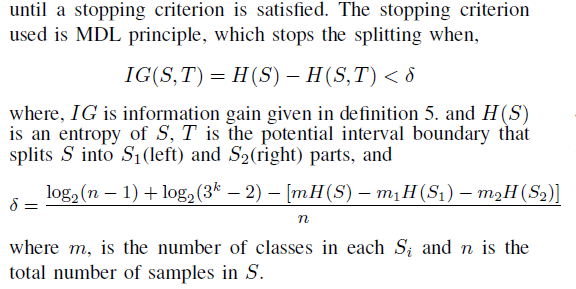
将连续属性离散化

Entropy-based discretization is a supervised discretization method, where the boundaries for discretization are selected by using class information entropy of candidate partitions. It considers one large interval containing all known values of an attribute then recursively partitions this interval into smaller sub-intervals. This recursive process stops until some stopping criterion, such as Minimum Description Length(MDL) principle [17] or an optimal number of intervals is achieved. Entropy is used as splitting criterion and MDL principle is used as stopping criterion. The Entropy is defined in definition.4.

基于熵的离散化是一种有监督离散化方法，在离散边界采用候选分类信息熵。它认为，一个大的区间包含所有已知的值的属性然后递归分割成更小的子区间的区间。这个递归过程直到停止准则达到时停止，如最小描述长度（MDL）原理[ 17 ]或区间的最优数量达到。熵作为分裂准则，MDL原则作为停止准则。熵的定义4定义。

这个熵和之前求信息增益一样。

于是就是信息增益作为分裂准则，MDL作为停止准则



**B算法（重点！！）**

前面是各种用到的字母的定义

算法是这样的：

3）把全部的隐私预算分给每一棵树P/B

4）把每一棵树的隐私预算分给每一层

5）对于每一棵树

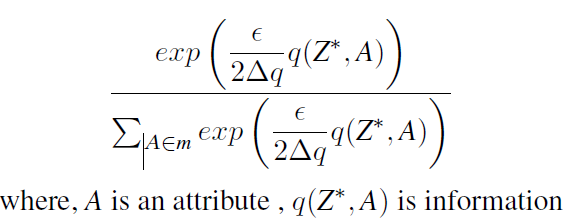
a）从D中有放回地选择τ个sample

b）生成每一棵随机森林中的树按照以下步骤，直到所有属性都用光或者d=0

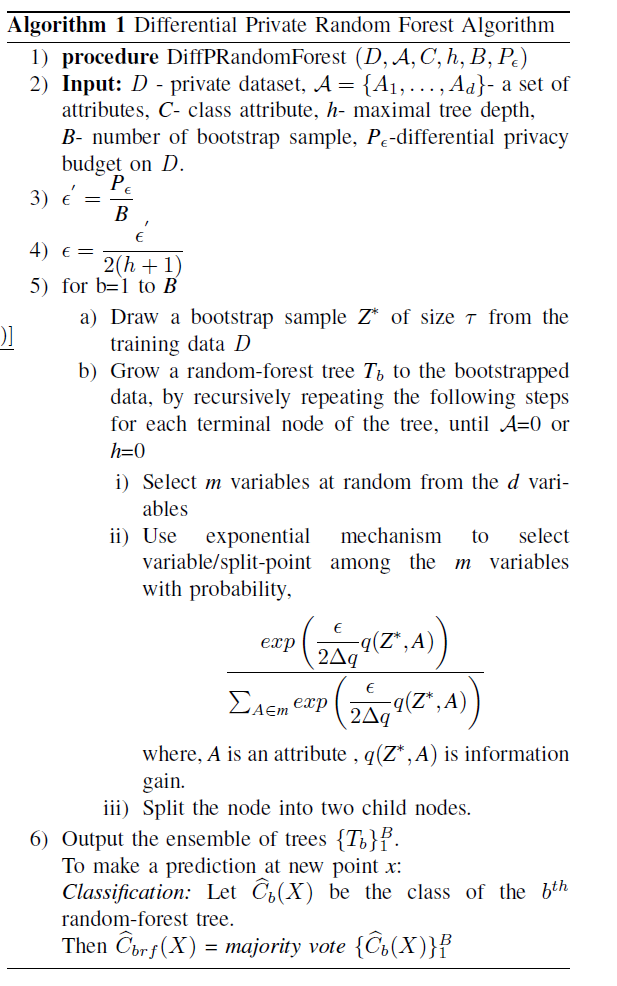
i）从A的d个属性中选择m个属性

ii）用指数机制从m个属性中选择分裂属性

iii）把这个node分裂成俩子node



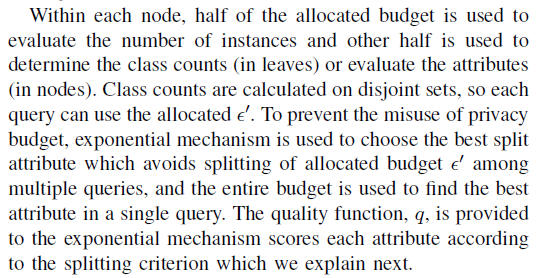
6）输出全部的树，对一个新的x来做预测



之后是隐私预算的分配。共Pϵ，每棵树是ϵ′，每一层是ϵ（树的深度是h），每一层每个node的隐私预算是ϵ，因为是不相交子集。

每一个node呢，分配一半的隐私预算用来估计实例的数量；另一半用来确定class count分类计数（leaves）或估计属性（nodes）。分类计数是在不相交子集上计数的，所以每一次查询可用的是ϵ′。

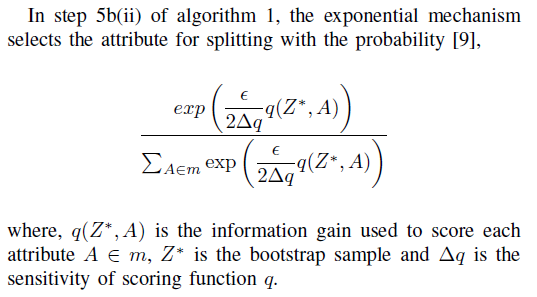
为了防止隐私预算的滥用，指数机制用来选择最佳分裂属性，整个预算用来选择最佳属性在一次查询中。



**C 分裂准则**

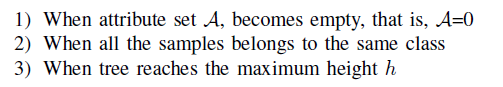
本文说了两种：InfoGain和max operator

让他们作为指数机制的打分函数，不同的打分函数有各自的敏感度



**D停止准则**

树递归停止有以下三个条件：



一旦某个条件达到了，带噪音的大多数类的统计来替代精确的类计数。

在DiffPID3中，如果叶结点中只有少量实例，噪声会覆盖真实计数，错误的class可能被选择，所以为了避免这个结果，引入一个阈值（依赖于加入的噪声），一旦实例数小于这个值，迭代就会停止。在我们的研究中，发现这儿阈值准则会降低分类准确度，且没用。

随机森林对任何大小的数据集都有很好的准确性，并且不论数据集有多少属性。这背后的主要原因是对数据的引导和随机选择变量。我们发现有没有阈值没啥影响。

**E 打分函数的敏感度**

InfoGain

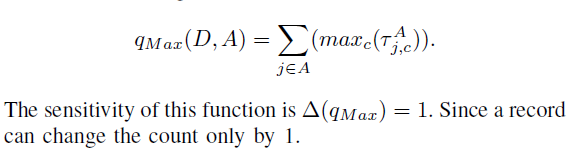


这个的敏感度是log2（C）C是类属性个数（如分成y or n C=2）



Max Operator

与节点错误分类率一致，选择最高出现频率的类。敏感度是1。



Gini指数的敏感度是2

我们的实验中，这三种的准确率对不同数据大小是相同的。原因是属性选择的随机性

**V EXPERIMENTAL RESULTS**

实验结果

分为合成数据和实际数据，连续和离散属性都有。

实验对不同隐私预算测了DIFFPID3算法的三种打分函数，DiffPRF的三种打分函数和随机森林本身的准确度。

结果看起来还蛮好的- -#

**VI. DISCUSSION**

这边主要讨论了，阈值限制没啥用，会降低准确度。

另外对不同的打分函数，它们自身的敏感度对准确性也没什么影响（原本那篇说InfoGain影响大，而MO影响小）