P3

data-based主要动机是通过只分析全集中的一个子集或者把全集转化为规模较小的表示

task-based主要动机是提高算法的时间，空间以及资源的效率

sampling:等概率的做一个投影，缺点是无法处理数据速率变化大的时间序列

load shedding:选取一些序列直接扔掉。缺点是可能扔掉一些至关重要的pattern

sketching:提取feature

aggregation:计算一些诸如平均值，方差之类的统计量，缺点是对数据分布严重倾斜的不适合

approximation:没啥好讲的，但要有error-bound

sliding-window:用户只关心数据流中最近一段的数据点

P5

Load Shedding主要用于监视系统，不断的对某一个滑动窗口内的数据做query，如果数据量过大，系统过载，就需要扔掉一些数据

P6

query plan中放置load shedder, shared-segment是Q5和Q6公用的operator

P7

F0计算很有技术含量，用O(1)的时间和空间

P8

K-median problem：n个点中找出k个中心点，使得其他点到最近的中心点距离和最小

只能扫描一遍，不能使用类似k-means的迭代法

Guha使用的空间,O(nk)的时间，并且证明了常量近似算法时间复杂度下界

Binary data streams clustering作者改进了传统的k-means方法来有效的处理二进制的data stream

High quality data stream clustering也是做k -median这个问题,quality的度量就是所有点到中心点的距离和

P12

主要罗列了data stream classification的research issues

1由于一遍扫描的要求，build classification 模型的速度要快于数据的速度

2 内存限制，需要一些sketching的技术

3 概念转移，随着数据不断的来，已经训练好的模型可能过时。可以使用一对分类器来投票解决

4 数据挖掘的结果的最终用户有可能是带宽有限的手机，传感器等终端。需要一个有效的表示方法，DFT

5 有时候用户并不关心数据流的数据，而是关心数据的变化，可以通过分类的结果变化来理解数据的变化

P13

frequent pattern mining可以用来挖掘关联规则

使用不同的窗口模型

1. 从某个开始点到当前点到时间段
2. 固定长度的时间窗
3. 给最近的数据点较高的权重