Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

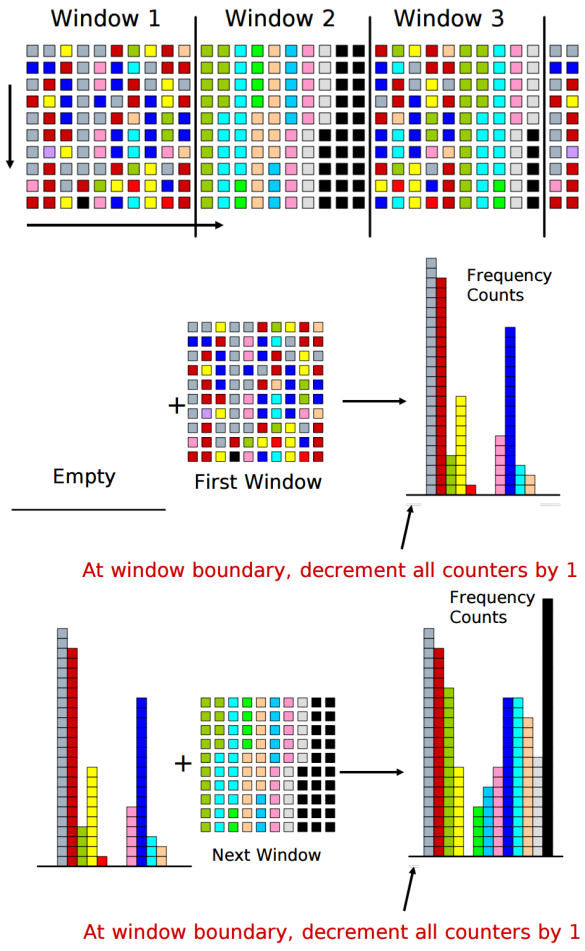
SS 2018

Exercise 6: Data Stream Clustering

Exercise 6-1 Lossy Counting

Before stream clustering, let's take a look at a more fundamental task in stream: count the occurrence of objects in a stream and output objects with a count larger or equal to some given threshold: $minSup \times L$, where L is the length of the stream up to now and $minSup$ is the given threshold (minimum support).

Lossy Counting is one of the basic algorithms that solve this problem. Given the windows size as $w = \frac{1}{\epsilon}$, the lossy counting algorithm works as the follows: cut stream into windows, process one window a time and prune histogram entries with 0 counts at each window boundary. The illustration is given below:



Please prove that

(a) the maximum count error (the maximum difference between the real count and the estimated count) of the lossy counting algorithm is ϵL

(b) the memory consumption, i.e., the number of entries stored in the histogram, is $O(\frac{1}{\epsilon}\log(\epsilon L))$. (Optional)

Exercise 6-2 Damped Window Model

Assume a special microcluster decaying mechanism, where all microclusters are fading out after each time stamp according to $f(t) = b^{-\lambda t}$. The weight of a microcluster is increased by 1 if it is updated (hit by a point in the current timestamp). Assume 1 data point per time stamp.

(a) What is the maximum weight of a microcluster?

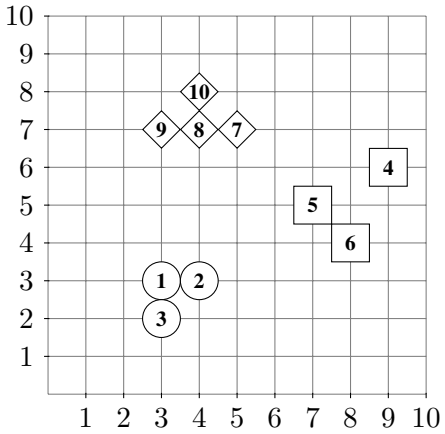
(b) What is the minimum time needed for a newly created microcluster to become potential (weight larger than τ)?

(c) What is the minimum time needed by a potential microcluster with weight w to became an outlier (weight less than τ)?

Exercise 6-3 Cluster Features

Given the following dataset:

ObjID	Cluster	X	Y	t
1	A	3	3	1.7
2	A	4	3	3.5
3	A	3	2	1.2
4	B	9	6	4.1
5	B	7	5	5.0
6	B	8	4	1.2
7	C	5	7	4.7
8	C	4	7	2.3
9	C	3	7	2.2
10	C	4	8	2.2



Compute the CluStream cluster features CFT for each of these three clusters.

A new observation in the stream is $p = (X = 8, Y = 5, t = 6.1)$.

Run the "online micro-cluster maintainance" of CluStream for this Point p.