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

Knowledge Discovery in Databases II

SS 2019

Exercise 8: Data Stream Clustering

Exercise 8-1 Change Detection: MONIC

MONIC is a change detection framework that does not assume a particular cluster model. Given the following dataset and set of clusters:

id	X	Y	t_0
1	0	0	0
2	0	1	0
3	2	0	0
4	2	1	0
5	1	0	1
6	1	1	1

where t_0 is the arriving time of the point. Assuming the aging function is $f(P,t) = 2^{-(t-P.t_0)}$.

The set of clusters ξ_0 at time $t = 0$ are $C_0^0 = \{1,2\}$ and $C_0^1 = \{3,4\}$. At time $t = 1$, ξ_1 contains only one cluster $C_1^0 = \{1,2,3,4,5,6\}$

Given $\tau = 0.75$, what external transitions can you detect here?

Exercise 8-2 Hoeffding trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1-2$ years, $2-7$ years, > 7 years)
- Gender (male, female)
- Residential area (urban, rural)

These are the first 8 examples.

Person	Time since license	Gender	Area	Risk class
1	$1-2$	m	urban	low
2	$2-7$	m	rural	high
3	>7	f	rural	low
4	$1-2$	f	rural	high
5	>7	m	rural	high
6	$1-2$	m	rural	high
7	$2-7$	f	urban	low
8	$2-7$	m	urban	low

- Incrementally construct a Hoeffding tree for this example.
  Use information gain and $\delta = 0.2$ and $N_{\text{min}} = 2$.
- Compute the value of δ at which the tree would still consist of the leaf only.