ADAPTIVE K-MEANS

A differenza del K-Means normale abbiamo uno stream di dati e un parametro C, ovvero il numero di istanze necessarie ad avviare la fase di inizializzione

FASE DI INIZIAZIONE

In questa fase valutiamo il miglior K per le C istanze che abbiamo. Valutiamo anche da quali K centroidi iniziare.

Cercare K

Stimare la densità di probabilità per ogni feature. Questa probability density function (PDF) avrà dei cambiamenti di direzione in termini di pendenza. Ogni coppia di cambiamenti consecutivi identifica una regione. Il numero di regioni è il nostro K.

Cercare i K centroidi

Dividere la PDF in K aree equiprobabili, e prendere il punto centrale come centroide.

Valutazione del cluster

Ogni feature porterà a diversi risultati in termini di K e centroidi iniziali. Valutare i migliori cluster tramite il silhouette coefficient.

CONTINUOS CLUSTERIG PHASE

Qui dobbiamo verificare se ci sono cambiamenti nello stream di dati tale per cui bisogna cambiare i cluster.
Lo fa tracciando la media e la deviazione standard dei dati in ingresso, e valutando un concept drift. Quando un concept drift viene valutato positivamente bisogna rifare la fase iniziale.

COMPLESSITÀ

1) Cercare K: $O(d \cdot e)$ con d numero di feature dato che vanno analizzati tutte le feature.

2) Runnare K-means: $O(e \cdot d \cdot K \cdot cs)$ dove cs è il numero degli insiemi di centroidi trovati.

3) Clusterizzare un nuovo dato $O(K)$

TOTALE: $O(K) + O(d \cdot e) + O(d \cdot e \cdot K \cdot cs)$

MUDI-STREAM

È sia density-based che grid-based.

CORE MINI CLUSTER

Sono dei feature vector speciali che contengono peso, centro, raggio e la distanza massima di un'istanza dalla media.

MUDI è diviso in due fasi: Online e Offline.

FASE ONLINE

In questa fase i CMC sono creati e tenuti aggiornati dai dati in streaming in ingresso.

Dato x dato corrente dello stream:

1) Trova il più vicino CMC rispetto a x: CMC_s

2) if CMC_s fitta bene con x
   aggiungi x a CMC_s.

   else
   Mappa x nella griglia
   if la griglia è abbastanza densa
   crea un nuovo CMC.

3) Quando arriva il periodo di pruning eliminare le griglie e i CMC più vecchi, ovvero con peso basso.

FASE OFFLINE

Usare i CMC per cercare i cluster veri.

1) Randomicamente scegliere un CMC non visitato. Marcarlo come visitato.

2) if il CMCp ha dei vicini
   Crea il cluster C.
   C.Add(CMCp)
   C.Add(vicinato(CMCp))
   for ∀ CMC in C
      Aggiungi il vicinato di CMC in C.
   else
      CMCp = rumore.

CONTRO MUDI

1) Non è buono per HD-Data, per la struttura a griglia