Institute for Informatics, Ludwig-Maximilians-Universität, München, Germany

Density-based Projected Clustering over High Dimensional Data Streams

Information Engineering and Computer Science Department University of Trento, Italy

High Dimensional & Dynamic Data:

- High dimensionality
  - overlapping / irrelevant/ locally relevant attributes
- Dynamic/ Stream data
  - only 1 look at the data (upon their arrival) / volatile data

! Clustering upon such kind of data is very challenging!!
  - both members and dimensions might evolve over time

☹ Most existing approaches deal with 1 aspect of the problem
☹ HPStream assumes a constant #clusters over time

Our Solution (HDDStream):

1st density-based projected clustering algorithm for clustering high dimensional data streams:

- High dimensionality ➔ projected clustering
- Stream data ➔ online summarize– offline cluster
- Density based clustering ➔ no assumption on #clusters/ invariant to outliers/ arbitrary shapes

☺ Quality improvement
☺ Bounded memory
☺ # summaries adapts to the underlying population

Projected MicroClusters:

For points $C=\{p_1, \ldots, p_n\}$ arriving at $t_1, \ldots, t_n$, the summary models both content and dimension preferences of C at t.

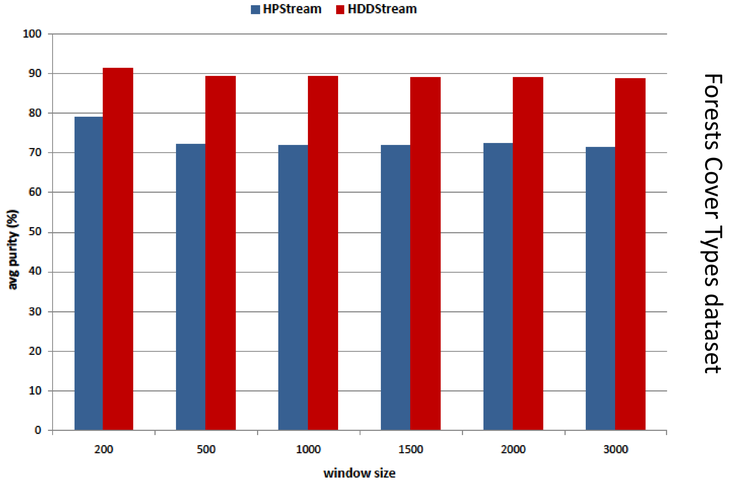
$mc \equiv mc(C,t) = <CF1(t), CF2(t), W(t)>$ — Content summary

- $CF1(t)$,/ $CF2(t)$: temporal weighted linear/square sum of the points
- $W(t)$: sum of points weights

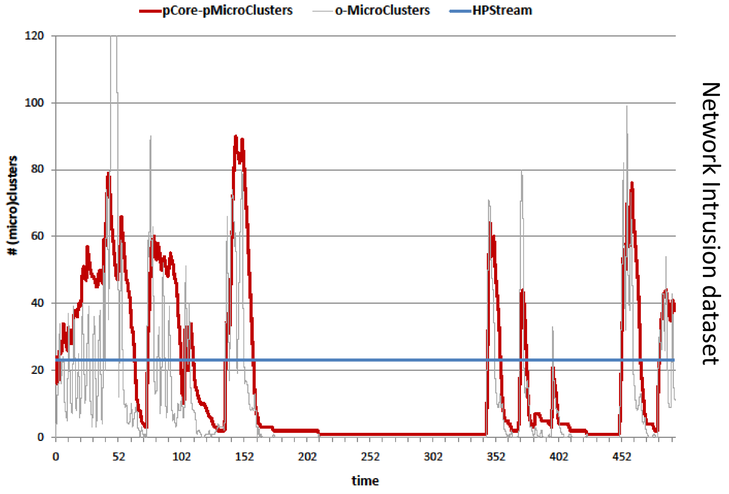
$\Phi(mc) = <\varphi_1, \varphi_2, .., \varphi_d>$ — Dimension preference vector

- j is a preferred dimension if: $VAR_j(mc) \leq \delta$

☺ Quality Improvement



☺ Change Detection & Monitoring



Clusters & Outliers

- Core projected microclusters (Core-pMC):
  - (1) $radius^{\Phi}(mc) \leq \varepsilon$ *(radius criterion)*
  - (2) $W(t) \geq \mu$ *(density criterion)*
  - (3) $P_{DIM}(mc) \leq \pi$ *(dimensionality criterion)*

The role of clusters and outliers in a stream often exchange:

- Potential core pMC (pCore-pMC):
  (1), (2) relax the density criterion $W(t) \geq \beta*\mu$, (3)
- Outlier MC (o-MC) :
  (1), (2) relax density criterion $W(t) \geq \beta*\mu$, (3) relax dim. criterion

Eirini Ntoutsi[1], Arthur Zimek[1], Themis Palpanas[2], Peer Kröger[1], Hans-Peter Kriegel[1]
[1]Institute for Informatics, Ludwig-Maximilians-Universität (LMU) München, Germany
[2]Information Engineering and Computer Science Department (DISI), University of Trento, Italy

Institute for Informatics, Ludwig-Maximilians-Universität, München, Germany

Density-based Projected Clustering over High Dimensional Data Streams

Information Engineering and Computer Science Department University of Trento, Italy

High Dimensional & Dynamic Data:

- High dimensionality
  - overlapping / irrelevant/ locally relevant attributes
- Dynamic/ Stream data
  - only 1 look at the data (upon their arrival) / volatile data

! Clustering upon such kind of data is very challenging!!
  - both members and dimensions might evolve over time

☹ Most existing approaches deal with 1 aspect of the problem
☹ HPStream assumes a constant #clusters over time

Our Solution (HDDStream):

1st density-based projected clustering algorithm for clustering high dimensional data streams:

- High dimensionality ➔ projected clustering
- Stream data ➔ online summarize– offline cluster
- Density based clustering ➔ no assumption on #clusters/ invariant to outliers/ arbitrary shapes

☺ Quality improvement
☺ Bounded memory
☺ # summaries adapts to the underlying population

Projected MicroClusters:

For points $C=\{p_1, \ldots, p_n\}$ arriving at $t_1, \ldots, t_n$, the summary models both content and dimension preferences of C at t.

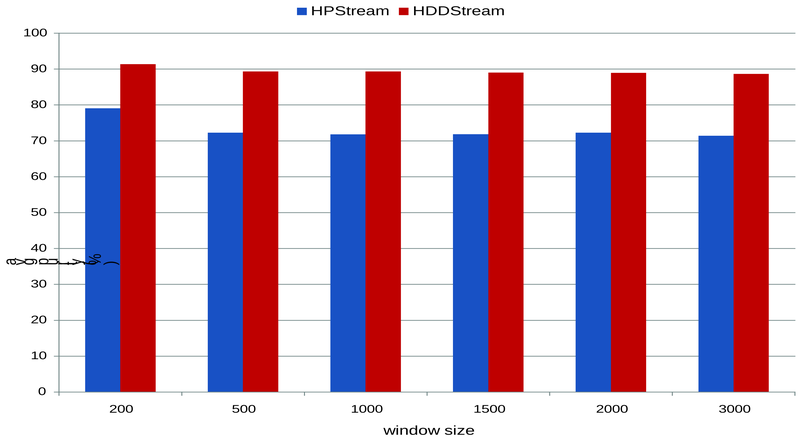
$mc \equiv mc(C,t) = \langle CF1(t), CF2(t), W(t) \rangle$ — Content summary

- $CF1(t)$,/ $CF2(t)$: temporal weighted linear/square sum of the points
- $W(t)$: sum of points weights

$\Phi(mc) = \langle \varphi_1, \varphi_2, .., \varphi_d \rangle$ — Dimension preference vector

- j is a preferred dimension if: $VAR_j(mc) \leq \delta$

☺ Quality Improvement



Forests Cover Types dataset

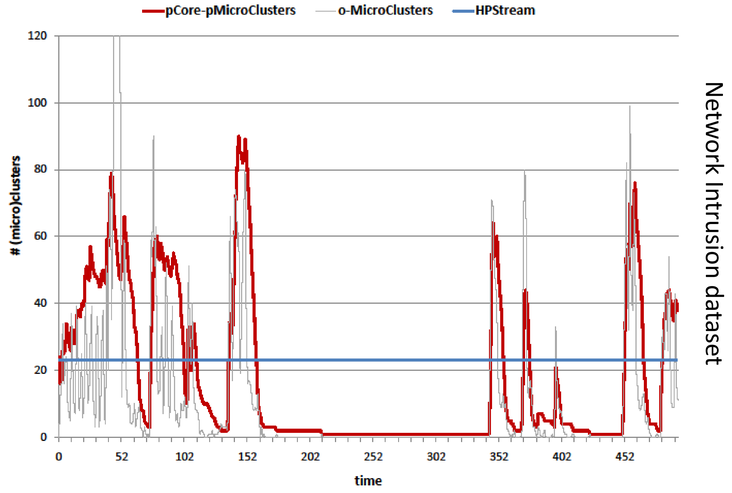
Clusters & Outliers

- Core projected microclusters (CORE-PMC):
  - (1) $radius^{\Phi}(mc) \leq \varepsilon$ *(radius criterion)*
  - (2) $W(t) \geq \mu$ *(density criterion)*
  - (3) $P_{DIM}(mc) \leq \pi$ *(dimensionality criterion)*

The role of clusters and outliers in a stream often exchange:

- Potential core PMC (PCORE-PMC):
  (1), (2) relax the density criterion $W(t) \geq \beta * \mu$, (3)
- Outlier MC (O-MC) :
  (1), (2) relax density criterion $W(t) \geq \beta * \mu$, (3) relax dim. criterion

☺ Change Detection & Monitoring



Eirini Ntoutsi[1], Arthur Zimek[1], Themis Palpanas[2], Peer Kröger[1], Hans-Peter Kriegel[1]
[1]Institute for Informatics, Ludwig-Maximilians-Universität (LMU) München, Germany
[2]Information Engineering and Computer Science Department (DISI), University of Trento, Italy