A Survey of Stream Clustering Algorithms

PAPER BY CHARU C. AGGARWAL
PRESENTED BY: DAVID PEREZ
What is a Stream?

- A sequence of data that can be read once (single-pass) or a small number of times (multiple-passes).
- Data may be transient (in the case of real time collection and processing).
Motivations and Issues

- Batch methods aren’t always possible – data is HUGE!
- Constantly evolving data.
- Space efficiency necessary in many cases.
- Scalability – We don’t know how large our data can get.
Data Stream

Data Abstraction Step (online component)

Clustering Step (offline component)

Data Partition

Data Structure for Statistic Summary

Time to time
Methods

- Partitioning Representatives
- Density Based Methods – Grid and Micro-Clustering Methods
- Probability Based Methods
Based on K-medians clustering

- Divide the whole data stream into chunks $D$, each with at most $m$ data points.
- Pick a set of $k$ points in each chunk, store it away as level-1 representatives.
- Repeat for each $D$.
- Once your representatives exceed your $m$, apply another level of clustering, using the stored weights in the process.
- Repeat this until done.
STREAM - Issues

- Not great at dealing with the evolving state of streams.
- How to determine optimal objective function value
CluStream

- Keeps microclusters of the data.
- Facilitates clustering through different time horizons.
- Uses Cluster Feature Vectors in clustering.
Cluster Feature Vector

- Structure that contains the following
- Sum of the squares of the data values
- Sum of the data values
- Sum of the squares of time stamps
- Sum of the time stamps
- The number of data item belonging to the cluster
- Point is, holds information that can be clustered.
Micro-Clusters

- Summary statistics about clusters.
- Think of “Snapshots” of data.
- Pyramidal way of storing data.
- Simplifies clustering huge amounts of data
Pyramidal Time Frame

- Information is stored at differing levels of granularity.
- Classified into orders: 1 to log(clock_time_elapsed).
- A snapshot of the i-th order occurs at times $a^i$ where $a$ is an integer.
- Snapshots only taken at the time when the clock value is exactly divisible by $a^i$.
- Redundancy in snapshots.

<table>
<thead>
<tr>
<th>Order of Snapshots</th>
<th>Clock Times (Last 5 Snapshots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>55 54 53 53 51</td>
</tr>
<tr>
<td>1</td>
<td>54 52 50 48 46</td>
</tr>
<tr>
<td>2</td>
<td>52 48 44 40 36</td>
</tr>
<tr>
<td>3</td>
<td>48 40 32 24 16</td>
</tr>
<tr>
<td>4</td>
<td>48 32 16</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
</tr>
</tbody>
</table>
WHO ARE YOU PEOPLE
DenStream

- Density based method using micro-clusters
- Clusters defined with a density threshold
- Different labels for different types of micro-clusters: potential core micro-cluster and outlier micro-cluster
- Labels of the micro-clusters can change accordingly
When new data arrives

- Try to insert into an existing p-micro-cluster
- Insert into an o-micro-cluster if possible
- If neither of these work, create a new o-micro-cluster
Questions?