Clustering by Synchronization

Christian Boehm, Claudia Plant, Junming Shao, Qinli Yang

Ludwig-Maximilians-Universität München
Munich, Germany
http://www.dbs.ifi.lmu.de
Email: shao@dbs.ifi.lmu.de
What is the Synchronization?

Synchronization: is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.
What is the Synchronization?

**Synchronization**: is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.

E.g. Opinion Formation
What is the Synchronization?

**Synchronization**: is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.

E.g. Opinion Formation

**Problem in a Group**
What is the Synchronization?

**Synchronization**: is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.

E.g. Opinion Formation

![Diagram](image-url)
What is the Synchronization?

**Synchronization**: is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.

**E.g. Opinion Formation**

![Diagram showing problem in a group, discussion, and synchronization with different opinions]

- Local
  - Opinion 1
  - Opinion 2
  - Opinion 3
**What is the Synchronization?**

**Synchronization:** is a phenomenon that a group of events spontaneously come into co-occurrence with a common rhythm, despite of the differences between individual rhythms of the events.

E.g. Opinion Formation

**Problem in a Group**  **Discussion**  **Synchronization**

![Diagram of Synchronization](image)

**Fig.:** Illustration of the process of Synchronization
Basic idea

- Each data object is regarded as a phase oscillator
Basic idea

- Each data object is regarded as a **phase oscillator**
- Each data object interacts with its similar objects through an **Extensive Kuramoto Model.**
Basic idea

- Each data object is regarded as a phase oscillator
- Each data object interacts with its similar objects through an Extensive Kuramoto Model.
- Investigate dynamic behaviors of objects over time.
Basic idea

- Each data object is regarded as a phase oscillator
- Each data object interacts with its similar objects through an Extensive Kuramoto Model.
- Investigate dynamic behaviors of objects over time.
  - Similar objects gradually synchronize together and form distinct clusters.
Basic idea

- Each data object is regarded as a **phase oscillator**
- Each data object interacts with its similar objects through an **Extensive Kuramoto Model**.
- Investigate dynamic behaviors of objects over time.
  - **Similar objects** gradually synchronize together and form distinct clusters.
  - **Outliers** tend to isolate from other objects and remain stable all the time.
Basic idea

C1

C2
Basic idea

Local Range

Extensive Kuramoto Model
Basic idea

Local Range  Extensive Kuramoto Model

C1  P1  P2  P3

C2
Basic idea

Local Range  Extensive Kuramoto Model
Key Points:

- How does each object interact with each other?
  - Cluster Model: Extensive Kuramoto Model

- How to determine the optimal range for local interaction?
  - Minimum Description Length
  - Kernel Density Estimation
Kuramoto Model

\[ \frac{d\theta_i}{dt} = \omega_i + \frac{K}{N} \sum_{j=1}^{N} \sin(\theta_j - \theta_i), \quad i = 1, \ldots, N \]

- where \( \omega_i \) describes the natural frequency, \( \theta_i \) is the phase of \( i \)-th oscillator and \( K \) is the coupling constant.

The global synchronization behavior rarely occurs in real-life systems. Local synchronization are more frequently observed.
Extensive Kuramoto Model

\[
\frac{dx_i}{dt} = \omega_i + \frac{S}{|Nb_\epsilon(x)|} \sum_{y \in Nb_\epsilon(x)} \sin(y_i - x_i)
\]

Let \( dt = \Delta t \), then:

\[
x_i(t+1) = x_i(t) + \Delta t \cdot \omega_i + \frac{\Delta t \cdot S}{|Nb_\epsilon(x(t))|} \cdot \sum_{y \in Nb_\epsilon(x(t))} \sin(y_i(t) - x_i(t))
\]
Let all objects have the same frequency $\omega$, the term $\Delta t \cdot \omega_i$ is the same for each object and thus ignored. $\Delta t \cdot S$ is a constant and simply fix it as 1.

**Model for Clustering**

$$x_i(t+1) = x_i(t) + \frac{1}{\left| Nb_\varepsilon(x(t)) \right|} \cdot \sum_{y \in Nb_\varepsilon(x(t))} \sin(y_i(t) - x_i(t))$$

- $i$-th dimensional phase of $x$ at time $(t+1)$
- Coupling Function
- Local Mutual Interaction
- Phase Diff.
Fully automatic clustering

- Optimal Local Range: Minimum Description Length

\[ L(D,M) = L(M) + L(D|M) \]

\[ \sum_{i=1}^{K} \sum_{j=1}^{|C_i|} \log_2 \left( \frac{N}{|C_i|} \right) + \sum_{i=1}^{K} \frac{p_i}{2} \log_2 (|C_i|) - \sum_{i=1}^{K} \sum_{x \in C_i} \log_2 (pdf(x)) \]

- Cluster-ID
- Free Parameters
- Data

Probability
Fully automatic clustering

Kernel density estimation

Fig.: Kernel density estimation vs. Gaussian estimation.
(a) Kernel density estimation (density and contour plot)
(b) Gaussian estimation (density and contour plot).
Experimental Evaluation

☑ Proof of Concept

Fig.: Performance of Sync w.r.t various aspects. (a) Arbitrarily shaped clusters (b) Clusters with various densities (c) Robust to outliers
Experimental Evaluation

Performance on Synthetic Data

- Without Data Distribution Assumption
- Fully automatically
Experimental Evaluation

- Sync VS. Classical Approaches

- K-Means (K=9)
- DBSCAN (ε=0.035)
- DBSCAN (ε=0.025)
- SC (K=9)
- Mean-Shift (b=6.3)
- Affinity Propagation (K=9)
Experimental Evaluation

- Sync VS. Parameter-Free Approaches

X-Means  RIC  OCI
Experimental Evaluation

- **Real Data - Wisconsin Data**

  The data set deriving from a study on breast cancer consists of 683 instances which are labeled to the classes malignant (M: 239 instances) and benign (B: 444 instances). Each instance is described by 9 numerical attributes.

**Performance:**

- Find the correct number of clusters;
- Discover almost all objects of each cluster with high recall (96.2% and 97.5%);
- All instances in each cluster match with corresponding type (with highest precision of 98.6% and 93.2%).
Experimental Evaluation

– Real Data - Wisconsin Data

Table 1. Performances on Wisconsin Data

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Sync</th>
<th>X-Means</th>
<th>RIC</th>
<th>OCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC [1]</td>
<td>0.154</td>
<td>0.183</td>
<td>0.182</td>
<td>0.154</td>
</tr>
<tr>
<td>NMI [2]</td>
<td>0.777</td>
<td>0.324</td>
<td>0.344</td>
<td>0.274</td>
</tr>
<tr>
<td>AMI [2]</td>
<td>0.777</td>
<td>0.322</td>
<td>0.343</td>
<td>0.272</td>
</tr>
<tr>
<td>AVI [2]</td>
<td>0.782</td>
<td>0.464</td>
<td>0.475</td>
<td>0.411</td>
</tr>
</tbody>
</table>


Experimental Evaluation

– Real Data - Diabetes Data

The diabetes data is a collection of medical diagnostic reports of 768 samples from a population. Each sample consists of eight significant risk factors which were chosen for forecasting the onset of diabetes.

Fig. 8: Illustration of the result of Sync on diabetes data.
Experimental Evaluation

– Real Data - *Diabetes Data*

Table 2. Performances on Diabetes Data

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Sync</th>
<th>X-Means</th>
<th>RIC</th>
<th>OCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>0.625</td>
<td>0.656</td>
<td>0.661</td>
<td>0.635</td>
</tr>
<tr>
<td>NMI</td>
<td>0.051</td>
<td>0.051</td>
<td>0.011</td>
<td>0.032</td>
</tr>
<tr>
<td>AMI</td>
<td>0.048</td>
<td>0.050</td>
<td>0.009</td>
<td>0.031</td>
</tr>
<tr>
<td>AVI</td>
<td>0.058</td>
<td>0.051</td>
<td>0.011</td>
<td>0.038</td>
</tr>
</tbody>
</table>
Desirable properties of Sync

- Novel clustering notion: Synchronization;
- Arbitrarily shaped clusters detection without data distribution assumption;
- Natural outlier handling;
- Fully automatic clustering in combination with MDL.
Thank you for your attention!