Data Particlization for Next Generation Data Mining

Takeaki Uno
National Institute of Informatics

Akihiro Yamamoto  Kyoto University
Yukinobu Hamuro  Kwansei Gakuin University
Kumiyo Nakakoji  Kyoto University

21/Apr/2016
Mining finds *structures* that are used by methods in upper layers from big, shallow meaning, sparse, noisy, and ill-granularity data.
**Data Abstraction**

**ex.) trajectory:** sequence of points $\rightarrow$ sequence of few places

- stayed in this area for 2 hours
- be here 10:40
- this street is common to many people passed through this part of street

**ex.) data points:**
- data objects $\rightarrow$
- local groups of similar objects
• Existing methods design “particle-like structures” independently
• Mining is not directed to good utilities of the methods
• Data particlization serves as the basis for the data analysis tasks
Machine Learning without Abstract

- Partition the data into two areas, including more reds, and not
- Even though attains high accuracy, the solution is “hard to understand” the mechanism
With Particles

Easier to get some meanings, or inspires

Here

Here
Why not Clustering?

Clustering finds (global) "classes", but particles are "structures"

... so, has many problems

huge small solutions, unbalanced sizes, skewed granularity
Why bad? ... because, the boundaries of the structures are not clear

The analogy: making the picture visually clearer  
sharpening edges, erasing noise, removing shadows, … and rearranging objects

At the same time, the accuracy in recognizing,  
classifying, and segmenting of the objects  
in the picture can be increased

Do the same in Bigdata!
A Proposed Method: Data Polishing

Reveal hidden structures by modifying the data based on feasible hypotheses

Graph clustering

Segmentation

Pattern mining

so that

• parts of the data are modified in such a way that any solution and structure would not be lost
• ambiguities are resolved, similar solutions are unified, and the number of solutions is reduced
• the quality of the data analysis will not be deteriorated
Preliminary Study for Graph Clustering

<table>
<thead>
<tr>
<th>the scale</th>
<th>original</th>
<th>polished</th>
</tr>
</thead>
<tbody>
<tr>
<td>#nodes</td>
<td>3,282</td>
<td>3,282</td>
</tr>
<tr>
<td>#edges</td>
<td>35,168</td>
<td>73,132</td>
</tr>
<tr>
<td>density</td>
<td>3.3‰</td>
<td>6.8‰</td>
</tr>
<tr>
<td>#cliques</td>
<td>32,953</td>
<td>343</td>
</tr>
</tbody>
</table>

Companies and their business relations

Prediction accuracy:
accuracy on customer attribute prediction by clustering methods

<table>
<thead>
<tr>
<th></th>
<th>clique</th>
<th>Newman</th>
<th>graph cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>60.60%</td>
<td>59.70%</td>
<td>60.03%</td>
</tr>
<tr>
<td>particle</td>
<td>71.36%</td>
<td>62.76%</td>
<td>67.78%</td>
</tr>
</tbody>
</table>

Noise robustness:
discovery rates of clusters (particles) by clustering methods

<table>
<thead>
<tr>
<th></th>
<th>polishing</th>
<th>Newman</th>
<th>graph cut</th>
</tr>
</thead>
<tbody>
<tr>
<td>noise 10%</td>
<td>100.00%</td>
<td>68.74%</td>
<td>76.10%</td>
</tr>
<tr>
<td>noise 40%</td>
<td>99.69%</td>
<td>7.91%</td>
<td>77.03%</td>
</tr>
</tbody>
</table>

+ acceptance ratio for dating proposal in marriage support: 13% → 29%
+ target size (users to show ads) without loss on internet advertisement: → 1/10
Extracting **needs** and **importance** from **data/user analysis**, and algorithms for **data polishing** and **semantic structures** of particles.