Anomaly Detection in Infrequently Occurred Patterns

Dong Wang
Principal Architect, wangdong13@baiduc.com
Baidu Inc.
Background

• Detailed explanation to a case mentioned in the talk

Agenda

• Introduction to Baidu
• What the problem is
• The idea and solution
• Results
Introduction to Baidu

• One of the largest search engines in the world
  ✓ Web/Image/Video/News/…

• Besides search, we also have
  ✓ Location Based Service – Maps
  ✓ Social/Knowledge – Tieba/Zhidao
  ✓ Online to Offline – Nuomi/Waimai
  ✓ Finance/Payment – Wallet
  ✓ Cloud computing – Cloud

• Covers more than 1 billion users in total
The problem – Anomaly Detection

• In theory it should not be a difficult problem ...
Example problem in Reality

• A metric’s curve around Spring Festival

• The results were lots of missed or false alarms
Some tried but failed Ideas

• Median compensation
• Time compensation
• Holt-Winters
• BP

• So we turn to some data mining methods.
The difficulties

• Infrequency means NO enough training data
  ✓ Holidays are infrequent

• We also cannot use the seasonality of time sequence
  ✓ In China, holiday dates are not fixed

<table>
<thead>
<tr>
<th>Year</th>
<th>Spring Festival</th>
<th>Dragon boat</th>
<th>Mid Autumn</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Feb. 19</td>
<td>Jun. 20</td>
<td>Sept. 27</td>
</tr>
<tr>
<td>2016</td>
<td>Feb. 8</td>
<td>Jun. 9</td>
<td>Sept. 15</td>
</tr>
<tr>
<td>2017</td>
<td>Jan. 28</td>
<td>May 30</td>
<td>Oct. 4</td>
</tr>
</tbody>
</table>
First Idea - Date Clustering

• Can we find as many as possible “similar” dates?
  ✓ Clustering on CDF of everyday’s data curve (k-means)
## More details

<table>
<thead>
<tr>
<th>Date</th>
<th>Green</th>
<th>Blue</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working day</td>
<td>208</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sat.</td>
<td>0</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>Sun.</td>
<td>0</td>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>Festivals</td>
<td>0</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>Specials</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

(3/1/2015 – 1/4/2016, Total: 310 days)
However, still some gap

- CDF reflects trend, not the exact values in points
Second Idea - Real time fixing

\[
\frac{\hat{X}(k)}{X(k-1)} = \frac{x(k)}{x(k-1)}
\]

\[
\frac{\hat{X}(k)}{\sum_{R(l)} X(l) + \sum_{W(l)} \hat{X}(L)} = \frac{x(k)}{\sum_{k-m+1} x(j)}
\]

\[
\frac{\hat{X}(k)}{\sum_{k-m+1} X(l)} = \frac{x(k)}{\sum_{k-m+1} x(j)}
\]
The experiment result
The real deployment

• Jan. 1st, 2017
Thanks for your Attention and any questions?

wangdong13@baidu.com