User response prediction in mobile advertising
Faustine Bousquet, Khanh Duong, Christian Lavergne, Sophie Lèbre, Anastasia Lieva

To cite this version:
Faustine Bousquet, Khanh Duong, Christian Lavergne, Sophie Lèbre, Anastasia Lieva. User response prediction in mobile advertising. ECML PKDD, Sep 2018, Dublin, Ireland. hal-02014821
Results of simulations

The objective of this design of experiment (DOE) is to evaluate limits of our EM algorithm when noise variance increases. Simulation settings on 700 campaigns whose CTR is simulated:

- Clusters are equidistributed
- Beta values are estimated on real experiments with $H = 5$ time slot and $S = 7$ day of week.
- Their absolute values vary from 0 to 18.13 with a median value equal to 0.18.

BIC/ICL estimated number of clusters VS simulated number of clusters

Comparison of estimated and simulated number of clusters when noise variance increases:

<table>
<thead>
<tr>
<th>Variance (median : 0.24)</th>
<th>Variance (median : 0.70)</th>
<th>Variance (median : 1.37)</th>
<th>Variance (median : 3.52)</th>
<th>Variance (median : 8.64)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>$2$</td>
<td>$3$</td>
<td>$4$</td>
<td>$5$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1$</td>
<td>$2$</td>
<td>$3$</td>
<td>$4$</td>
<td>$5$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1$</td>
<td>$2$</td>
<td>$3$</td>
<td>$4$</td>
<td>$5$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1$</td>
<td>$2$</td>
<td>$3$</td>
<td>$4$</td>
<td>$5$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$1$</td>
<td>$2$</td>
<td>$3$</td>
<td>$4$</td>
<td>$5$</td>
</tr>
<tr>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

Some estimation:
- If we take the second piechart whose variance median is 0.75:
  - When the number of cluster simulated is $K=2$, the number of cluster estimated is $K=2$.
  - When the number of cluster simulated is $K=13$, the number of cluster estimated is $K=9$.

Campsions confusion matrix (K=4):

First Clustering result

First results are on CTR metric. We worked with 700 campaigns which started and ended between May the 10th and July the 10th. Our model included 2 temporal variables:
- day of week (cardinality $S = 7$)
- time of the day into buckets (cardinality $H = 5$)

Optimal number of clusters by BIC/ICL criteria:

Inferred profiles:
- Beta values are very different from one cluster to another.
- Same observation about clusters size : they include from 9 to 123 campaigns.
- Time slot and day of week effect seem to be significant.

Cluster 1 (nc = 120) Cluster 2 (nc = 173) Cluster 3 (nc = 123) Cluster 4 (nc = 45)
Cluster 5 (nc = 35) Cluster 6 (nc = 42) Cluster 7 (nc = 16) Cluster 8 (nc = 19)
Cluster 9 (nc = 33) Cluster 10 (nc = 9) Cluster 11 (nc = 37) Cluster 12 (nc = 39)
Cluster 13 (nc = 9)

Conclusions

- First results are encouraging. We obtain different cluster profiles with only few temporal variables. Evaluating algorithm, its limit and statistical results still remain in progress and are part of the challenges of this thesis. Also, we should work with domain expert to validate pertinence of clusters.
- Context varies on ads (size, type) and device (OS, model and so on) will enrich the model. Hopefully, these new variables and their interactions will lead to more and more homogeneous clusters.
- Once the clustering objective achieved, the goal will be to determine an adaptive predictive model for each cluster. The model should be able to scale the large amount of request received per second (around 1 million).

References