What can pre-release search traffic profiles tell us?

Oliver Schaer
Nikolaos Kourentzes
Robert Fildes

40th ISMS Marketing Science Conference
15th of June 2018
Clustering for new product forecasting

Applications

- **Model parameter clustering** (Goodwin et al. 2013, Bayus, 1993)
- **Time series shape clustering** (Basallo-Triana et al. 2017)
- **Product features clustering** (Hu et al. 2018, Baardman et al. 2017)
- **Functional clustering** (Sood et al. 2009)

For pre-launch forecasting additional information needed, i.e. expert judgment

- **Judgmental bias** (Belvedere & Goodwin, 2017; Tyebjee, 1987)
- **Consumer preferences change during pre-launch phase** (Meeran et al. 2017)
The literature reports positive findings when using online buzz for pre-launch forecasting (e.g. Kim and Hanssens 2017; Xiong & Bharadwaj 2014; Kulkarni et al. 2012)

- Predominately regression type models

Why clustering?

- Restricting into clusters increases homogeneity
- More robust to rely on cluster profile means rather individual forecasts?
- Regression based models are not readily applicable to obtain competitor product forecasts
Suggested approach

Create Clusters
- Generate features from product characteristics, sales information and search traffic

Train classifier
- Classify search traffic and known product features against obtained cluster solutions

Predict new product
- Obtain the cluster profile of sales by allocating a new product to their respective clusters
Empirical evaluation

Dataset

- Global physical video game sales of 240 popular video games from VGChartz
- IGN Score and MetaCritic information
- Weekly Google Trends data with game title as keyword

Split data into training and test set (70% / 30%)

Objective is to assess the performance of cluster profiles for predicting the opening week and total life-cycle sales.
Obtaining cluster profiles

Features

• Product characteristics (publisher, genre, no. release, month)
• Sales information (actuals and time for life-cycle stages, Gompertz model parameters)
• Pre-release search (sum of search as soon signal is available, trend line, bass model parameters)

Estimation

• Mixed-data converted into binaries using Euclidean distances (and weighted product features)
• Various clustering algorithms including k-means, k-medoids, hierarchies and hyperplane cutting
k-means cluster solution
Some insights on cluster profiles (k-means)

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (n = 30)</th>
<th>Cluster 2 (n = 32)</th>
<th>Cluster 3 (n = 12)</th>
<th>Cluster 4 (n = 3)</th>
<th>Cluster 5 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(First week sales)</td>
<td>11.370</td>
<td>12.016</td>
<td>15.358</td>
<td>16.326</td>
<td>12.787</td>
</tr>
</tbody>
</table>
Some insights on cluster profiles

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (n = 30)</th>
<th>Cluster 2 (n = 32)</th>
<th>Cluster 3 (n = 12)</th>
<th>Cluster 4 (n = 3)</th>
<th>Cluster 5 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(First week sales)</td>
<td>11.370</td>
<td>12.016</td>
<td>15.358</td>
<td>16.326</td>
<td>12.787</td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Weeks until sales 25%</td>
<td>4.000</td>
<td>3.000</td>
<td>1.000</td>
<td>1.000</td>
<td>2.000</td>
</tr>
<tr>
<td>Weeks until sales 95%</td>
<td>44.500</td>
<td>34.500</td>
<td>19.000</td>
<td>14.000</td>
<td>24.000</td>
</tr>
</tbody>
</table>
## Median cluster profiles

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (n = 30)</th>
<th>Cluster 2 (n = 32)</th>
<th>Cluster 3 (n = 12)</th>
<th>Cluster 4 (n = 3)</th>
<th>Cluster 5 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(First week sales)</td>
<td>11.370</td>
<td>12.016</td>
<td>15.358</td>
<td>16.326</td>
<td>12.787</td>
</tr>
<tr>
<td>Weeks until sales 25%</td>
<td>4.000</td>
<td>3.000</td>
<td>1.000</td>
<td>1.000</td>
<td>2.000</td>
</tr>
<tr>
<td>Weeks until sales 95%</td>
<td>44.500</td>
<td>34.500</td>
<td>19.000</td>
<td>14.000</td>
<td>24.000</td>
</tr>
<tr>
<td>First signal</td>
<td>24.000</td>
<td>38.500</td>
<td>40.000</td>
<td>40.000</td>
<td>40.000</td>
</tr>
<tr>
<td>GT trend slope</td>
<td>2.421</td>
<td>5.864</td>
<td>118.284</td>
<td>200.755</td>
<td>10.779</td>
</tr>
</tbody>
</table>
Some insights on cluster profiles

Median cluster profiles

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (n = 30)</th>
<th>Cluster 2 (n = 32)</th>
<th>Cluster 3 (n = 12)</th>
<th>Cluster 4 (n = 3)</th>
<th>Cluster 5 (n = 91)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(First week sales)</td>
<td>11.370</td>
<td>12.016</td>
<td>15.358</td>
<td>16.326</td>
<td>12.787</td>
</tr>
<tr>
<td>Weeks until sales 25%</td>
<td>4.000</td>
<td>3.000</td>
<td>1.000</td>
<td>1.000</td>
<td>2.000</td>
</tr>
<tr>
<td>Weeks until sales 95%</td>
<td>44.500</td>
<td>34.500</td>
<td>19.000</td>
<td>14.000</td>
<td>24.000</td>
</tr>
<tr>
<td>First signal</td>
<td>24.000</td>
<td>38.500</td>
<td>40.000</td>
<td>40.000</td>
<td>40.000</td>
</tr>
<tr>
<td>GT trend slope</td>
<td>2.421</td>
<td>5.864</td>
<td>118.284</td>
<td>200.755</td>
<td>10.779</td>
</tr>
<tr>
<td>IGN Score</td>
<td>7.975</td>
<td>7.917</td>
<td>8.779</td>
<td>8.000</td>
<td>8.250</td>
</tr>
</tbody>
</table>
Classification task

Random Forest to learn search traffic and product features against the obtained cluster solutions

Forecast exercise

• Benchmark against regression based Random Forest (same inputs on opening week and total life-cycle sales)
• Compare the MAE against the mean cluster profile
Forecasting process

**Clustering**
- **Inputs**
  - Sales information
  - Google Trends information
  - Product characteristics
- **Output**
  - K number of clusters

**Classification**
- **Inputs**
  - Google Trends information
  - Product characteristics
- **Output**
  - Predicted cluster allocation

**Candidate: Clustering**
- **Mean of cluster profile**
- **Benchmark: Regression**
  - Product level forecast
  - MAE of opening week sales and total life-cycle sales
### Overall forecasting performance

#### Relative Mean Absolute Error of clustering over regression

<table>
<thead>
<tr>
<th></th>
<th>k-means</th>
<th>k-medoids</th>
<th>Hierarchical</th>
<th>Hyperplanes</th>
</tr>
</thead>
<tbody>
<tr>
<td>First week sales</td>
<td>1.141</td>
<td>1.178</td>
<td><strong>0.896</strong></td>
<td>1.709</td>
</tr>
<tr>
<td>Total life-cycle sales</td>
<td><strong>0.768</strong></td>
<td>0.890</td>
<td>1.750</td>
<td>1.210</td>
</tr>
</tbody>
</table>

Number of clusters, $k = 5$

All cluster inputs without diffusion model parameters
Conclusion

Modelling implications

• Extending the new product clustering literature by considering pre-launch search profiles.
• Outperforms more common regression based approach

Managerial implications

• More intuitive view on the drivers of the cluster profiles
• With known clusters any new product can be assessed, also those of our competitors
Next steps

Optimise the clustering via cross-validation for

- selecting the best clustering algorithm;
- find optimal number of clusters;
- obtain distance metric settings.

→ to improve the opening week performance

Forecasting competitor success only training clusters with company internal sales information

Assess different lead times of Google Trends
Thank you!

o.schaer@lancaster.ac.uk
@oliverschaer
www.lancaster.ac.uk/lums/cmaf


Appendix: cluster profile of sales

First week sales

Total sales

log(opensales) vs Clusters

log(tot.sales) vs Clusters