Applications of Unsupervised Learning to Business Meta-Data, using Yelp Data

Eric Wang *Stanford University, Dept. of Statistics*
Charles Zhang *Stanford University, Dept. of Statistics*

MOTIVATION

- Physical businesses (restaurants, hotels, retail) provide a variety of "products".
- QUESTION: Can we group these businesses by their "products"?

- Businesses choose advantageous locations to sell their products.
- QUESTION: Can we distinguish between different types of business locations and business groupings?

DATA

- We use data from Yelp and apply unsupervised learning techniques to this business meta-data to answer these questions.
- 61K Businesses, 10 Cities, 1.5M reviews, 500K tips.

SOM Business Neighborhoods

Example Business Attribute Clusters

Seafood / Steakhouse	*Fast Tex-Mex*	*Night Life*
• Center Cut Steakhouse	• Taco Bell	• Diamond Club
• Fuji Japanese Steakhouse	• El Pollo Loco	• Baccarat Lounge
• Mastro's Ocean Club	• Chipotle	• Artisan Hotel Boutique

Methodology/Procedure

1. Form business "neighborhood" using SOM.
2. Pre-process business meta-data using PCA
3. Choose the correct k for K-Medoids using the Gap Statistic,
4. Cluster restaurants by their attributes using K-Medoids.
5. Assign businesses to their neighborhood with their class.
6. Cluster the neighborhoods using K-Means.

Algorithms

K-Medoids is computationally more expensive, but more robust to outliers and categorical data thank K-Means.

Algorithm 1 K-Medoids

for $j = 1 : k$ **do**
  $m_i \leftarrow x_{random}$ ▷ Randomly initialize medoids
end for
while $\Delta \sum_i ||x_p - x_i||_2^2 > \gamma$ **do**
  for $i = 1 : m$ **do**
    $c_i \leftarrow arg\min_j ||x_i - m_j||_2$ ▷ Assignment Step
  end for
  for $j = 1 : k$ **do**
    $m_j \leftarrow arg \min_{x_i : class_i = j} \sum_{x_p : c_p = j} ||x_p - x_i||_2^2$ ▷ Maximization Step
  end for
end while

SOM is an online algorithm that preferences equal sized clusters and uses a uniform initialization.

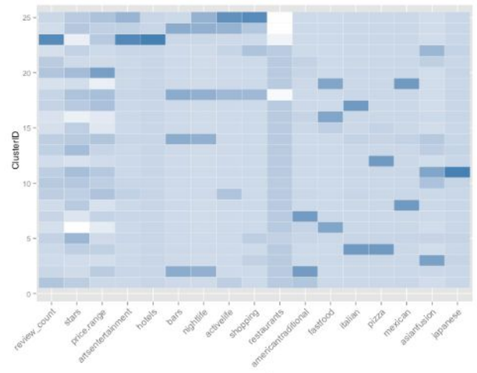
Algorithm 2 Self-Organizing Maps (Online)

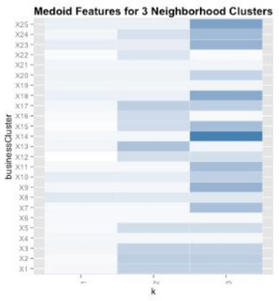
repeat ▷ Repeatedly present the training data to the algorithm
  for $i = 1 : m$ **do**
    $w_i \leftarrow arg\min_{n_j} ||x_i - n_j||_2^2$ ▷ Find the closest neuron
    for $n_j \in w_i \cup \Gamma(w_j)$ **do**
      $n_j \leftarrow n_j + \alpha\Theta(n_j, w_i)(x_i - n_j)$ ▷ Update the closest neuron, and its neighbors
    end for
  end for

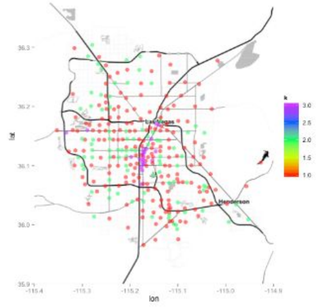
Gap Statistic compares the actual within-cluster error to the expected within-cluster error and provides a decision rule for choosing the value of *k*.

$$Gap(k) = \frac{1}{B}\sum_b log(W_{kb}^*) - log(W_k)$$

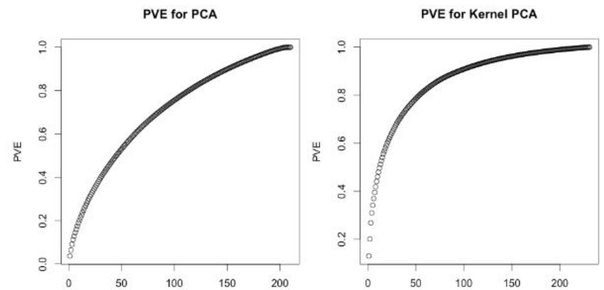
Examining Business Clusters and Relation to Neighborhoods

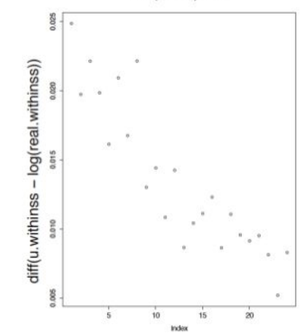


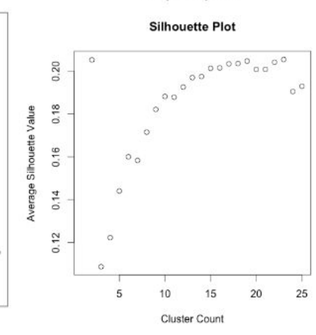




Dimension Reduction and Evaluation of Cluster Count







Discussion

- We find that businesses cluster across attributes beyond the coarse categories provided in Yelp meta-data
- Further, the interpretability of our results improved significantly by adjusting scaling and using k-medoids
- Clustering neighborhoods seems to be dominated by volume of user reviews; seem to identify neighborhoods with high activity
- *Extensions include:*
  - Assigning clusters to other cities and comparing mix of businesses between cities
  - Further exploring relationship between neighborhoods and business mix
  - Replacing SOM with GSOM
  - Clustering businesses by users and their ratings with dimension reduction. This could then be contrasted with the business attribute clusters