



CLUSTERING A CUSTOMER BASE USING TWITTER DATA

VANESSA FRIEDEMANN



I. INTRODUCTION

There are numerous applications within the retail industry for clustering large populations including:

1. allows company to target advertising
2. identifies untapped markets for expansion
3. feeds recommendation systems

Our approach uses publicly available Twitter data to perform customer clustering for a chosen company, Nike.

Why Twitter Data?

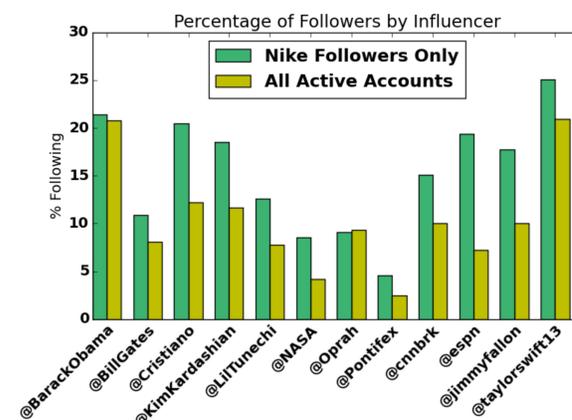
Large-scale private-sector data, such as sales history and loyalty account information, is prohibitively difficult to obtain. For any company in need of information regarding customers other than its own, there is a need for an alternative.

II. DATA SET AND FEATURES

Features

For each sample user, the final feature set includes:

1. statuses posted
2. number of followers
3. number of accounts following
4. language
5. influencer following relationships



Pre-processing Steps

Requirements of k-means lead us to perform the following pre-processing steps:

1. distance correlated with similarity -> map language to geographic coordinates
2. non-categorical features -> PCA on influencer following relationships
3. equally-weighted features -> feature range normalization

In addition, we discard features that exhibit very low variance.

Tools Used

To gather the data from Twitter, we use Tweepy. We use SQLite to store the data in a database. For all machine learning algorithms, we use the scikit-learn Python package.

III. METHODS

Silhouette Coefficient

We compute the optimal number of clusters for k-means by selecting the first k value with a silhouette coefficient larger than $\Gamma = 0.7$. The silhouette coefficient is computed as

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & , a(i) < b(i) \\ 0 & , a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & , a(i) > b(i) \end{cases} \in [-1, 1]$$

This is the ratio of the similarity of a sample to its own cluster versus the next best cluster.

Clustering Performance

Given the optimal value for k chosen above, we evaluate the clustering performance on our data set by computing the following

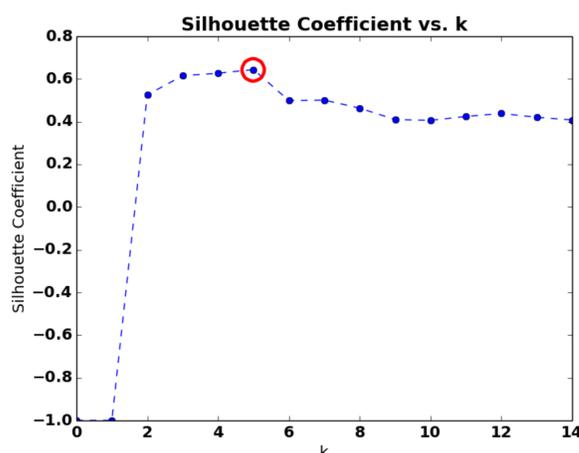
$$q(x, k) = \frac{\sum_{i=1}^k \left(\frac{\sum_{j=1}^m 1\{x^{(j)} \in c_i\} \|x^{(j)} - c_i\|_2}{\sum_{j=1}^m 1\{x^{(j)} \in c_i\}} \right)}{\sum_{i=1}^k \left(\frac{\sum_{j=1}^k 1\{i \neq j\} \|c_i - c_j\|_2}{\sum_{j=1}^k 1\{i \neq j\}} \right)}$$

This is the ratio of average intra-cluster variation to average inter-cluster distance.

IV. RESULTS

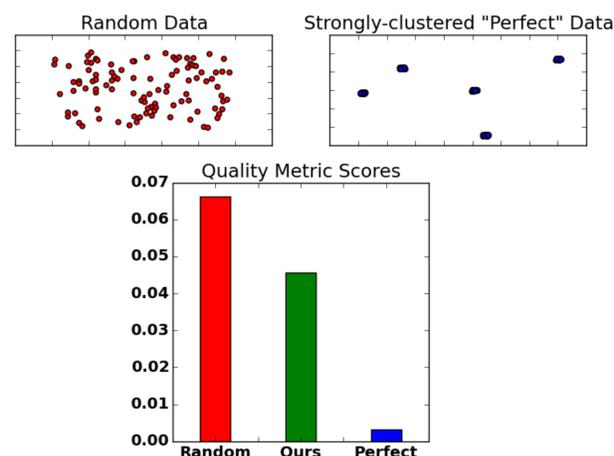
Selecting the Number of Clusters

We bound the number of clusters between $k = [2 : 15]$. We plot the silhouette coefficient vs. number of clusters. The below figure shows that the optimal number of clusters for our data set is $k = 5$.



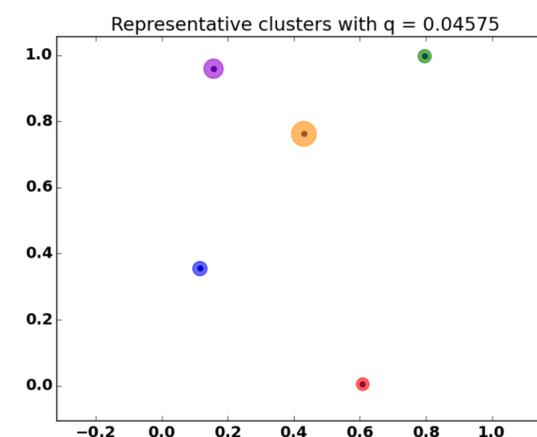
Measuring Clustering Performance

We compute a q score using $k = 5$ for randomized data, our data set, and perfectly clustered data. Low values of q correspond to better clustering performance. The plot below indicates that our clustering performance is in between that of ideal data and that of randomized data.



Visualizing Clusters

Since our data is in R^{13} , we cannot readily visualize it. Instead we show a representation of a clustering in R^2 that achieves the same q score. This suggests that our own data set can be cleanly clustered in its higher dimensionality space.



V. CONCLUSION

Discussion

We have shown that the quality of clustering achievable using our techniques is midway in between that of ideal and randomized data. Further, the clusters created are remarkably pronounced and well-defined. The outcome of these experiments further reinforces the credibility of using social media data for market analysis.

Future Work

Future research on this topic should focus on applying these algorithms to a more reliable data set. More factual data sources might include credit card transactions, newsletter subscriptions, and Amazon shopping cart history.