User Segmentation Models

Data Science Institute
COLUMBIA UNIVERSITY

Meiqi Chi, TaeYoung Choi, Aditya Huddedar, Leshen Sui, Siwen Tang
Industry Mentors: Charles Menguy, Handong Zhao
Faculty Mentor: Prof. Eleni Drinea

Data Science Capstone Project with Adobe Inc.

User Segmentation Using Characteristic Traits

We present a set of diverse algorithms that automatically create the most optimal segments of the users based on their behavioral traits. The Adobe dataset contains binary information about approximately 137 million user profiles and 5197 traits. The sparsity of the data is about 99.82%. The trait "800" is ubiquitous and thus contributes the most to the long tail effect.

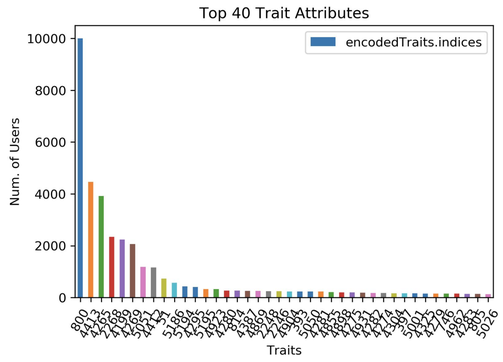

Figure 1. Top 40 Trait Histogram.

Auto-encoder + K-means

We use an auto-encoder to learn a lower dimensional non-linear transformation of our dataset. Since the dataset is binary, we use a Binary Cross Entropy Loss in the auto-encoder. We then run a K-means algorithm on the encodings obtained. In order to compare clusterings with different K's, we use a Calinski-Harabasz Index customized to effectively handle binary data.

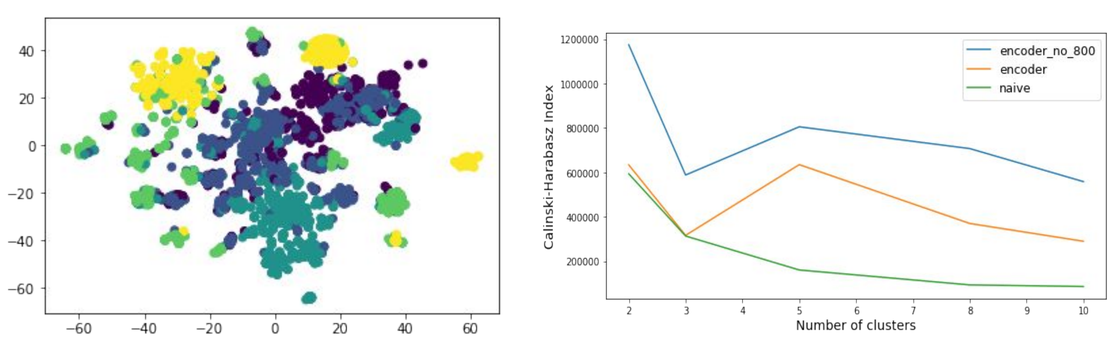

Figure 2. Auto-encoder based clustering results. K=5 achieves the best result.

Matrix Factorization + K-means

We build an Implicit ALS to decompose the original sparse matrix into low dimensional user factors and trait factors, and then apply K-means clustering on top of user factors. In Figure 3, the left side plot is based on minimizing the MSE on all observed entries equally, and the right side plot is built by weighing each trait with the reciprocal of its relative frequency.

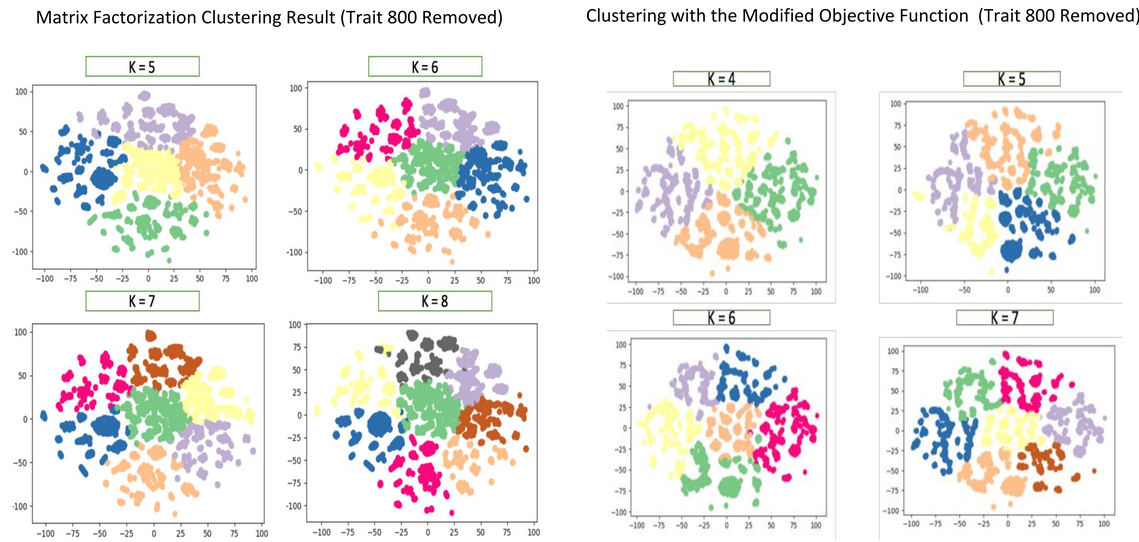

Figure 3. Matrix Factorization based clustering results.

Conclusions

Dimensionality reduction is the key to sparse binary data clustering. Even though the dataset has more than the 5000 traits per user, the optimal number of user segments is very small (around 5).

Acknowledgments

We would like to thank Prof. Eleni Drinea for giving us the opportunity to work on this project. We would also like to thank Charles and Handong for guiding us along the path.

References

1. Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis.
2. Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering.