Explainable Deep Clustering on National Survey of Tax and Benefit Data(NaSTaB) using Autoencoder and Dimension Reduction Techniques

Bonwoo Koo

Collaborated with
Insu Choi (KAIST ISysE Ph.D Candidate), Woosung Koh (Yonsei Economics B.S Candidate)

Financial Engineering Lab

KAIST

Motivation

(출처: 마이데이터 종합포털)

(출처: 코스콤뉴스룸)

- Korea's **'MyData'** initiative promotes 'Open Finance' and 'Open Banking', allowing major financial institutions and digital platforms to consolidate consumer data into a single hub with the individual's consent (Park et al., 2021).
- Although a sophisticated deep clustering framework has been recently introduced, it is essential to provide empirical or theoretical justification for the selection of specific algorithms within the framework.
- Specifically, challenges persist in **managing high-dimensional financial consumer datasets** and in **conveying the integration of AI to those without expertise in the field**.
- Aim to **systematically identify optimal clustering components** based on the dataset, with an **emphasis on interpretability and visualization** throughout our comprehensive customer data mining system.

Data Description

- High-dimensional cross-sectional data from the Korea Institute of Public Finance (KIPF). A total of 8,798 individual households and 14,837 unique individuals are sampled. A total of 238 features were selected from the household data set, and a total of 286 features were chosen from the individual data set. In short, the household data set, H∈R^(8,798×238), and the individuals' data set, H∈R^(14,837×286).

Methodology

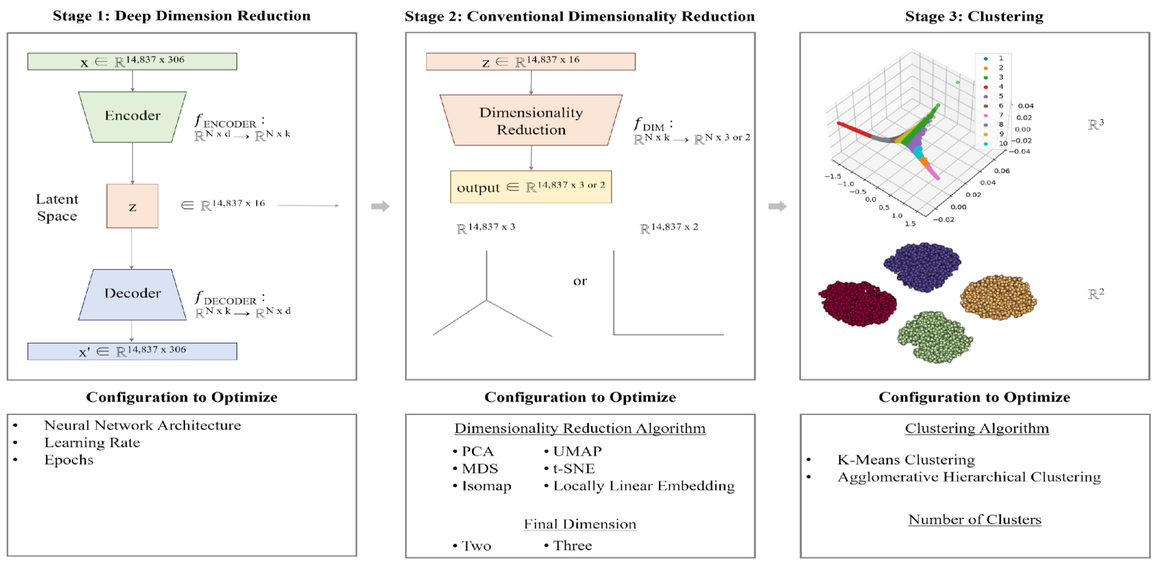


Figure 1. Proposed Three-Stage Clustering Method

- Stage 1: Autoencoder

$$f_{ENC}: R^{N \times d} \rightarrow R^{N \times k}$$
$$f_{DEC}: R^{N \times k} \rightarrow R^{N \times d}$$
$$l_{AE} = \| X - f_{DEC}(f_{ENC}(X)) \|_F^2$$

where N: number of samples, d: original dimension of data, k: dimension of latent vector

- Stage 2: Dimensionality Reduction
  - PCA
    - A classical dimensionality reduction technique that linearly transforms the high-dimensional data into a new coordinate system where the variation in the data can be described with fewer dimensions than the initial data
  - Manifold Learning(MDS, HLLE, Isomap, t-SNE, UMAP)
    - Non-linear dimensionality reduction technique that seeks to describe the data as low-dimensional manifolds embedded in high-dimensional spaces
- Stage 3: Clustering
  - K-means Clustering
    - $arg\ min_s \sum_{i=1}^{k} |S_i| * Var(S_i)$
    - when S_i is the cluster set, $\{S_1, S_2, \ldots, S_k\}$, given a set of observations $\{x_1, x_2, \ldots, x_n\}$ and the mean μ_i of points in S_i and $Var(S_i)$ is the squared Euclidean distances between the mean and the data point x
  - Hierarchical Clustering (Agglomerative)
    - Hyperparameters: metric, distance
      - Metrics: Euclidean and Manhattan distance
      - Linkage: Single, Complete, Average, and Ward
- Shapley Additive explanation (SHAP)
  - **To identify the most influential features within the high-dimensional feature set in the clustering process**. Based on the significant features, clusters are interpreted and customer segmentations are profiled.
  - By employing game theory, we can obtain the mean of the estimated Shapley values by averaging the conditional expected values for each data column.
  - Shapley value $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$
    - N : the set of total input variable, S : the set except i-th variable in the total input variable, $v(S)$: the contribution that the remaining subset except the i-th data contributed to the result, $f_x(S \cup \{i\})$: the total contribution including the i-th data.

Experiment Design

- First, the hyperparameter of the autoencoder is optimized. Then, all possible configurations of stages two and three, which consists of different dimensionality reduction and clustering algorithm, are experimented to **find the optimal configuration with highest performance ranking score**.
- Performance metrics for aggregated scaled ranking score:
  - Silhouette: $s(x_i) = \frac{min\{d_l(x_i)\} - a(x_i)}{max\{min\{d_l(x_i)\}, a(x_i)\}}$
    - where x_i : element in cluster π_k, $a(x_i)$: average distance of x_i to all other elements in the cluster π_k
  - Calinski-Harabasz score: $\frac{trace(BCSM)}{trace(WCSM)} * \frac{N-k}{k-1}$
    - where BCSM is a between-cluster scatter matrix, WCSM is a within-cluster scatter matrix, N is the number of points in the data set, and k is the number of clusters
  - Davies-Bouldin score: $\frac{1}{k}\sum_{i=1}^{k} R_i$ where $R_i = \max R_{ij}, i \neq j$, $R_{ij} = (S_i + S_j)/M_j$.
    - S_i is the sum of the average distances from each point in cluster i to the centroid of its cluster, and M_{ij} is the distance between the two cluster centers.
  - Equally weighted aggregation score: $\sum_{s \in S} MinMax(s \mid C_2, \ C_3)$
    - C_2 and C_3 refer to the sets of configurations to optimize in Stage 2 and 3, respectively. Score set, S := {Silhouette, Calinski–Harabasz, 1 – Davies–Bouldin}.

Results & Analysis

- Feature Engineering – Processed Data Visualization

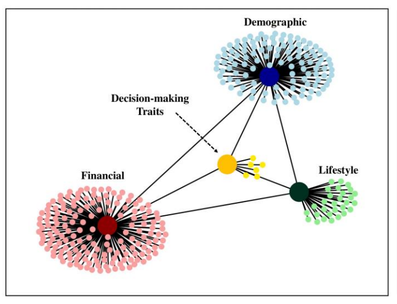


Figure 2. All Features' Category Network

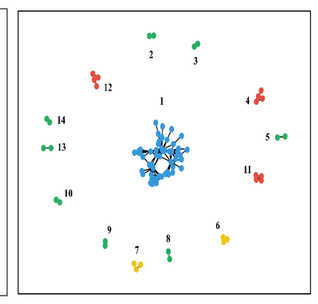


Figure 3. Numerical Features' Relationship Cluster Network

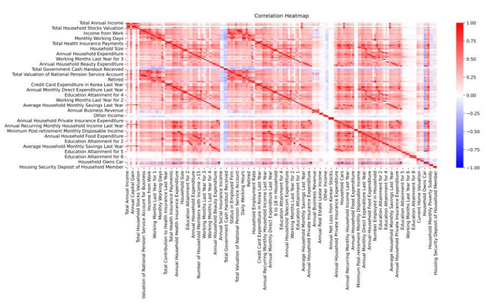


Figure 4. Numerical Features' Relationship Heatmap

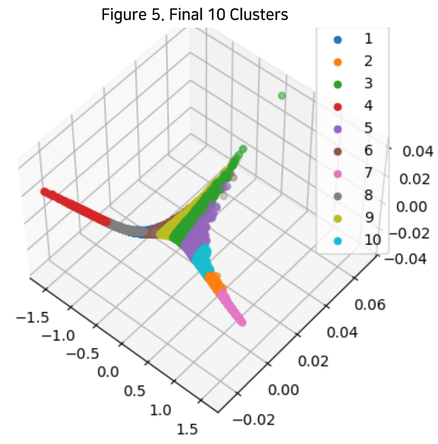
- Before three-stage clustering, two unique networks are created to represent the data and feature set intuitively.
  - All 306 features in a three-dimensional space via feature categorization (set C = **{Financial, Demographic, Lifestyle, Decision-making Traits}**).
- Figure 3 and Figure 4 are illustrations of our relationship cluster network and the commonly used relationship heatmap, for features of numerical data type. The proposed network in Figure 3 is a supplementary visualization method to the commonly used relationship heatmap and matrix.

- Experiment Results

Methodology						Scaled Score			
Stage 1	Middle Dimension	Stage 2	Final Dimension	Stage 3	Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin	Aggregated Score
Autoencoder	**16**	**Isomap**	**3**	**K-means**	**10**	**0.3971**	**1.0000**	**0.4191**	**1.0000**
Autoencoder	16	Isomap	2	K-means	10	0.3983	0.9974	0.4188	0.9994
Autoencoder	16	PCA	2	K-means	10	0.3982	0.9963	0.4188	0.9992
Autoencoder	16	PCA	3	K-means	10	0.3950	0.9970	0.4173	0.9992
Autoencoder	16	MDS	3	K-means	10	0.3781	0.9765	0.3950	0.9929

Table 1. Top 5 Scaled Scores

Figure 5. Final 10 Clusters



- In total, 1944 clustering experiments have been conducted.
- The optimal three-stage clustering configuration is Linear **Autoencoder** to extract the latent vector of data in 16-dimensional space, intermediate stage dimensionality reduction by **Isometric Mapping** to reduce into the third dimension, bringing the data in **3-dimensional space**. Finally, **K-means clustering** is used to define the final **10 clusters**. The final clusters, in 3 dimensions, are visualized in Figure 5.
- Across 13 different Stage 1 and 2 configurations and the baseline of clustering the original data with 305th dimension in the absence of Stage 1 and 2, all types of Stage 1 and 2 configurations yielded higher scores than 0.6048 which is the mean score of the baseline.
- The three-stage clustering algorithms were able to discern differences between clusters more effectively **by maintaining the latent features and gradually reducing the data dimensionality through autoencoder and dimensionality reduction technique**.

- SHAP
  - Identify the most impactful features of each cluster in this multi-stage clustering process, aiming to uncover the heterogeneity across the clusters.

Feature	Data Type	Category	Factor Affecting Investment
Number Employed in Household	Numerical	Demographic	Occupation
Regular Working Hours	Categorical	Demographic	Occupation
Employed Firm's Headcount	Numerical	Demographic	Occupation
Annual Recurring Monthly Household Income Last Year	Numerical	Financial	Income
Joined Public Pension	Categorical	Decision-making Traits	Others
Status in Employed Firm	Numerical	Demographic	Occupation
Regularity of Working Hours	Numerical	Demographic	Occupation
Monthly Pre-tax Income	Numerical	Financial	Income
Household Housing Security Deposit	Numerical	Financial	Asset
Annual Household Expenditure	Numerical	Financial	Expenditure

Table 2. Significant Features and Description

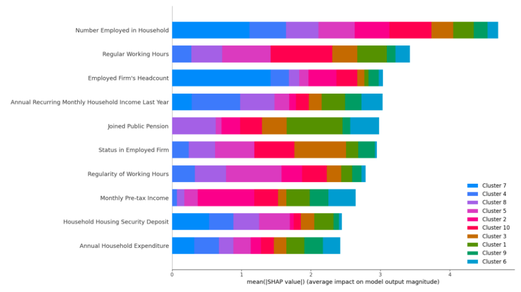


Figure 6. Significant Features Ranking by SHAP Values

- Customer Profiling

Cluster	Discretionary Labeling: Profile
1	Typical Upper-middle Class
2	Non-affiliated Lower Class
3	Middle Class with Government Aid
4	Wealthiest
5	Frugal Middle Class
6	Self-employed Hard-Working Middle Class
7	Highly Risk-averse Poorest
8	Typical Upper Class
9	Typical Middle Class
10	High-expenditure Lower Class

Table 3. Cluster-based Customer Segment Labeling

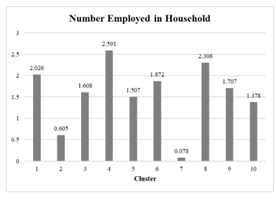


Figure 7. Sample Heterogeneity of Clusters

Figure 8. Sample Radial Chart of Top 10 Features

- Figure 7 illustrates the heterogeneity of clusters across features by keeping the feature constant on each bar graph. Next, potential customer segment groups, as represented by clusters, are profiled via the top ten features in and are visually summarized via radial charts in Figure 8.

Limitations

- Single Dataset
  - Future studies could investigate the generalizability of our approach to datasets such as text, images, and multi-modal datasets, and develop innovative methods to optimize the configuration search process
- No comparison with other deep clustering methodology
  - Further research can explore alternative dimensionality reduction and clustering methods and evaluating their effects on clustering performance and downstream applications

Conclusion

- We propose an **end-to-end system with network-based EDA and explainable three-stage clustering approach** to provide a more detailed understanding of high-dimensional data using numerous visualization approaches.
- Allow **systematic search for optimal configuration** of the three-stage clustering framework and provide **interpretability and visualization** factors in the three-stage clustering process
- Our incorporation of SHAP values in customer profiling reflects an ongoing effort to provide businesses with tools that aid in-firm and out-firm stakeholders like **non-technical managers and regulators in understanding black-box-like computational systems**.
- Spanning **from socioeconomics to business intelligence**, our proposed method can be integrated depending on the specific domain and problem set

References

- Park, J. K., Park, S. K., & Lee, B. G. (2021). Priority of Challenges for Activation of Mydata Business: K-Mydata Case. Ksii Transactions on Internet & Information Systems, 15(10).
- Wang, F., & Sun, J. (2015). Survey on Distance Metric Learning and Dimensionality Reduction in Data Mining. Data Mining and Knowledge Discovery, 29(2), 534-564.