On Not-Supervised Models

Ana Sofia Gutierrez Tejada
Universidad EAFIT
Medellin, Colombia
asgutierrt@eafit.edu.co

Olga Lucía Quintero Montoya
Universidad EAFIT
Medellin, Colombia
oquinte1@eafit.edu.co

***Abstract*—The following delves into an extensive exploration of unsupervised clustering methodologies, conducted on a Billionaires characterization dataset. The investigation evaluates distance-based approaches, revealing the inherent limitations of solely relying on distance metrics in clustering procedures. The project proceeded to examine the implications of incorporating density based algorithms and Neighborhood based clustering in a single pipeline. Building upon this, a comprehensive exploration of dimensionality manipulation both into lower-dimensional spaces through advanced techniques such as UMAP and into higher-dimensional spaces through Autoencoder Multi-Layer Perceptrons. The study culminated in an examination of the performance of the state-of-the-art HDBSCAN algorithm, notwithstanding the complexity of the dataset. Through this multifaceted investigation, a comprehensive understanding of the intricacies and nuances involved in diverse clustering methodologies within the context of complex data analysis was attained.**

***Index Terms*—Autoencoder, UMAP, Density-Based Clustering, K-Nearest Neighbors clustering, Non-Supervised Clustering.**

I. Introduction

We look to explore diverse unsupervised clustering methodologies, focusing on the analysis of a comprehensive Billionaires characterization dataset [1]. The dataset provided encompasses diverse information concerning different billionaires, including their final net worth, age, Consumer Price Index (CPI) changes within their respective countries, Gross Domestic Product (GDP) statistics of the countries, total tax rates prevalent in the countries, and the population sizes of the nations. This comprehensive dataset enables an analysis of the factors influencing the wealth accumulation and distribution among the billionaires, Table I offers insights into the economic, demographic, and policy contexts that shape their financial standing as either self-made (68%) or not (32%). The critical examination of the limitations inherent in distance-based approaches, although conceptually straightforward, provide a foundational understanding of distance-based clustering techniques and serve as a starting point for more sophisticated and refined clustering algorithms [2].

After shedding light on the necessity for more nuanced clustering strategies, delving into the implications of integrating density-based algorithms and Neighborhood based clustering within a unified pipeline. Additionally, the project undertakes a thorough investigation into the manipulation of dimensionality, both in lower-dimensional spaces through advanced methodologies like UMAP, and in higher-dimensional spaces leveraging Autoencoder Multi-Layer Perceptrons.

	Final Worth	Age	CPI	GDP	Tax rate
count	2640	2575	2456	2476	2458
mean	4623.79	65	4.364169	1.2E+13	43.96
std	9834.24	13.258	3.623763	9.6E+12	12.15
min	1000	18	-1.9	3.2E+09	9.9
25%	1500	56	1.7	1.74E+12	36.6
50%	2300	65	2.9	1.99E+13	41.2
75%	4200	75	7.5	2.14E+13	59.1
max	211000	101	53.5	2.14E+13	106.3
scale	e+03	e+01	e+00	e+13	e+01
empty	0	65	184	164	182

TABLE I: Exploratory analysis of Billionaires Dataset

Density-based clustering algorithms offer effective solutions for identifying clusters within complex datasets based on the density distribution of data points. These algorithms operate by exploring the density characteristics of data points, thereby delineating clusters as regions of higher density separated by areas of lower density. Leveraging the notion of density-based clustering, Mountain clustering, as a hierarchical algorithm, aims to discern clusters characterized by diverse shapes and densities in intricate datasets. This approach constructs a hierarchical cluster representation by iteratively merging clusters, relying on density and proximity criteria. In contrast, the Subtractive Density Cluster method represents a simpler density-based clustering technique, identifying clusters through the detection of density peaks within circular neighborhoods surrounding individual data points. While demonstrating efficiency in datasets exhibiting well-defined clusters, this method exhibits limitations in effectively handling datasets characterized by irregular shapes or varying densities. The utilization of such density-based algorithms contributes significantly to the comprehensive understanding and extraction of complex patterns within the Billionaires dataset.

Finally, Internal validation indices are integral tools in the evaluation and comparison of the performance of clustering algorithms, enabling the quantitative assessment of the quality and efficacy of clustering solutions. These indices serve as metrics for gauging the compactness, separation, and overall coherence of the clusters generated by different algorithms. The selection of an appropriate internal validation index is crucial in ensuring the accuracy and reliability of the evaluation process. Commonly used indices, such as the Davies-Bouldin index, Dunn index, and Calinski-Harabasz index, offer distinct perspectives on cluster quality, considering factors such as cluster separation, compactness, and the ratio of between-

cluster dispersion to within-cluster dispersion. The comparative analysis of these indices involves assessing their numerical values, where lower values of certain indices, such as Davies-Bouldin, indicate better clustering, while higher values of other indices, such as Dunn and Calinski-Harabasz, suggest improved cluster separability and coherence. The study finally involves a detailed evaluation of the performance of the cutting-edge HDBSCAN algorithm against the previously mentioned.

II. Methodology

A. Data Manipulation

The original domain of the data points was expanded and contracted using a combination of an Autoencoder Multi-Layer Perceptron (MLP) configuration and the UMAP (Uniform Manifold Approximation and Projection) embedding technique. Autoencoders are neural network architectures designed for unsupervised learning tasks, with the primary objective of encoding and decoding high-dimensional data efficiently. The implementation consist of an encoder network that compresses the input data into a lower-dimensional representation (latent space) and a decoder network that reconstructs the original data from this representation.

On the other hand, UMAP is a dimensionality reduction and manifold learning technique that specializes in preserving the global structure and topology of data while reducing dimensionality. Unlike traditional methods like Principal Component Analysis (PCA) or t-SNE, UMAP leverages graph theory and optimization techniques to create a low-dimensional representation of data that retains local and global patterns effectively [3]. UMAP is particularly valuable for visualizing complex data structures and enhancing the interpretability of high-dimensional datasets. This expanded and contracted transformations not only aid visualization but also facilitate the discovery of hidden patterns and relationships within the data.

B. Clustering Algorithms

1) Naive methods: The two initial exploratory methods discussed in the context of distance-based clustering represent naive implementations of concepts introduced in the classroom setting. Both algorithms typically commence by selecting an arbitrary data point from the dataset and subsequently establishing a cluster with this point as its centroid. This clustering process is executed in one of two approaches: firstly, by including data points located within a specified distance threshold from the initial point, thus forming a group; secondly, by selecting the k nearest neighbors of the initial point, thereby constituting a cluster. Subsequently, the algorithm proceeds to choose another data point, which serves as a reference for the creation of another cluster. This iterative process continues until all data points within the dataset have been allocated to clusters.

2) K-Nearest Neighbors clustering: The subsequent distance-based algorithms implemented in this study encompass the K Nearest Neighbors (KNN) method and its Fuzzy Clustering modification. KNN is a straightforward clustering technique that assigns each data point to the cluster of its closest centroid. On the other hand, the Fuzzy Clustering modification of KNN introduces a degree of membership for each data point to a cluster, proportional to its distance to each centroid. This modification allows data points to belong partially to multiple clusters, with a degree of membership parameter representing the level of association.

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij}^{m} \, ||x_j - c_i||^2 \tag{1}$$

Both algorithms evaluate Eq. 1 on each cluster from 1 to k and each point from 1 to n to stop its iterations. u_ij is 1 if point j belongs to cluster i and zero otherwise. In general, each cluster is formed around a centroid, that is iteratively defined as its mean data point. In the Fuzzy CNN algorithm, this mean is weighted proportionally to the degree of pertinence of each point to the cluster [4].

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}$$

Finally, as proposed by Torra [5], a value of $m = 2$ will be used for the Fuzzy classification through our experiments on each of the distance metrics defined [6].

3) Density-based clustering: Mountain clustering is a hierarchical density-based clustering algorithm. The core idea is to identify clusters whose density is relatively high compared to the surrounding regions. This method starts by calculating the density between data points using 2 weighted by a density radius factor, as shown in figure 1. By adding the density of each data point with respect the others, the algorithm assigns a local density value to each point, and chooses the highest evaluated to serve as cluster centers. It iteratively clusters the data points that are within a clustering radius of these centroids in order of magnitude. The clustering radius should be somewhat greater than the density radius in order to ensure the identification of cluster centers that are sufficiently separated.

The mountain clustering algorithm assumes that points on a uniformly constructed grid over the domain of the dataset are the possible centroids for the clusters, however the substractive density method assumes the dataset points as possible clusters. This hierarchical approach allows both algorithms to capture clusters of different scales and densities, making it robust in handling datasets with irregularly shaped or overlapping clusters.

$$F(xi, xj) = e^{-\frac{1}{4r_a^2}||xi-xj||^2} \tag{2}$$

4) Literature found clustering algorithm HDBSCAN: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a popular density-based clustering

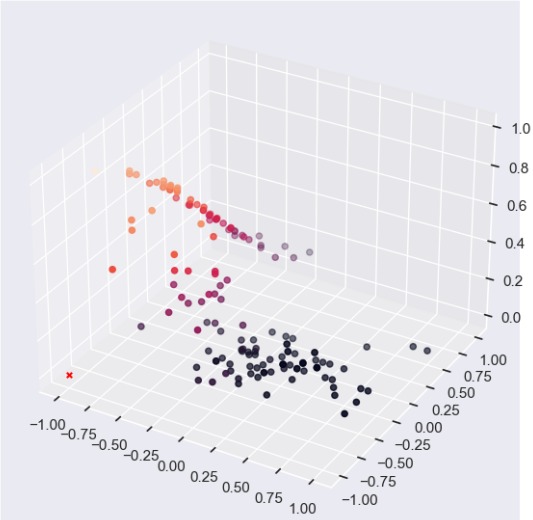


Fig. 1: 2D Density evaluation around $\hat{x} = [-1, -1]$ of Iris dataset

algorithm known for its capability to find clusters of varying shapes and sizes, as well as its ability to identify noise and outliers in a dataset. It constructs a hierarchical representation of the input data, allowing it to find clusters of varying densities. By analyzing the density of points, it differentiates between noise, clusters, and outliers, providing a flexible approach for clustering complex and large datasets. HDBSCAN is widely used in various fields such as data mining, pattern recognition, and image analysis, where the identification of clusters of different densities is essential. The Scikit-Learn implementation is used as a last step comparison of this classifier against the proposed pipeline [7].

C. Clustering methods evaluation

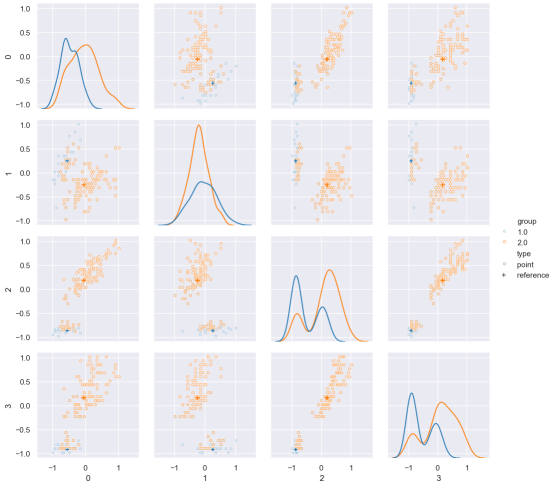
Internal validation indices, such as the Davies-Bouldin index (DBI), Dunn index, and Calinski-Harabasz index (CHI), play a crucial role in assessing the quality of clustering algorithms in data analysis [8]. The Davies-Bouldin index, proposed by David L. Davies and Donald W. Bouldin, measures the compactness and separation between clusters by considering both the intra-cluster similarity and inter-cluster differences. It computes the average similarity between clusters while also taking into account the distances between cluster centroids. A lower DBI value suggests better clustering, indicating well-separated and compact clusters.

On the other hand, the Dunn index, formulated by J. C. Dunn, evaluates the compactness and separation of clusters based on the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. A higher Dunn index signifies well-separated and internally coherent clusters. It is particularly useful for identifying compact and well-separated clusters within a dataset.

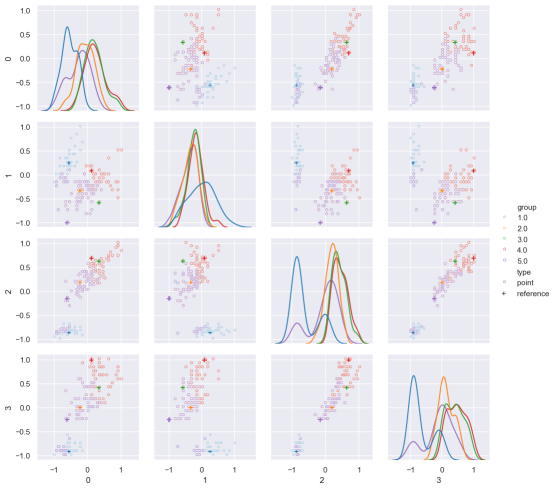
Finally, the Calinski-Harabasz index, developed by T. Calinski and J. Harabasz, assesses the ratio of the between-cluster dispersion to the within-cluster dispersion. A higher CHI value indicates better-defined and more separated clusters. This index is valuable for evaluating the homogeneity and separation of clusters, especially when the number of clusters is not predetermined. These indices are used for accurately evaluating the performance of clustering algorithms and are the base of the proposed unsupervised classification pipeline.

III. Results

In the context of distance-based approaches, the necessity of supplementary parameters becomes evident. Figure 2 illustrates that the exclusive reliance on distance often leads to the emergence of clusters that significantly overlap. This occurs due to the absence of classification criteria beyond the hierarchical arrangement solely based on the proximity to a particular point. As a consequence, the limited consideration of additional variables and their associated influences renders the clustering process vulnerable to the production of less distinct and more ambiguous clusters.



(a) Unsurpervised classification based on distance thresholds



(b) Unsurpervised classification based on a fixed number of neighbors

Fig. 2: Naive distance-based classifications on Iris Dataset

Moreover, as illustrated in Figure 5, the internal validation indices were computed across various clusters resulting from the implementation of K-NN and C-NN classification. The algorithms were parameterized with the determined number of clusters identified by the density classifiers. The assessment entailed 8 different radius values and 5 distance metrics for both the subtractive and mountain clustering algorithms. The extreme index evaluation values are presented on Table II.

The methodology was replicated using data encoded within a higher-dimensional space. With the implementation of an Autoencoder Multi-layer Perceptron encoding into 7 characteristics as opposed to the initial 5. The internal validation indices are presented in Figure 6, demonstrating improved performance across all indices. Although the expansion of the feature space offers additional dimensions for differentiation, determining the optimal number of clusters remains challenging. This predicament arises from the complexity associated with segregating distinct categories of billionaires based solely on the provided dataset. At the same time, Table III presents the index for the best and worst clusters made on the encoded data.

Conclusively, the process was applied to data embedded within a reduced-dimensional space utilizing the Uniform Manifold Approximation and Projection (UMAP) technique, resulting in a 2-dimensional feature representation. The internal validation indices are visually depicted in Figure 7, while the extremities of the index values are tabulated in Table IV. Notably, the cluster identified as the most optimal in accordance with the Calinski-Harabasz index is graphically depicted in Figure 3. Additionally, representations of one of the exemplary clusters derived from the original dataset and the dataset transformed through the multi-layer perceptron (MLP) encoding, are respectively presented on Appendix FIgures 8 and 9, emphasizing the structural characteristics and distinguishing features of the identified clusters.

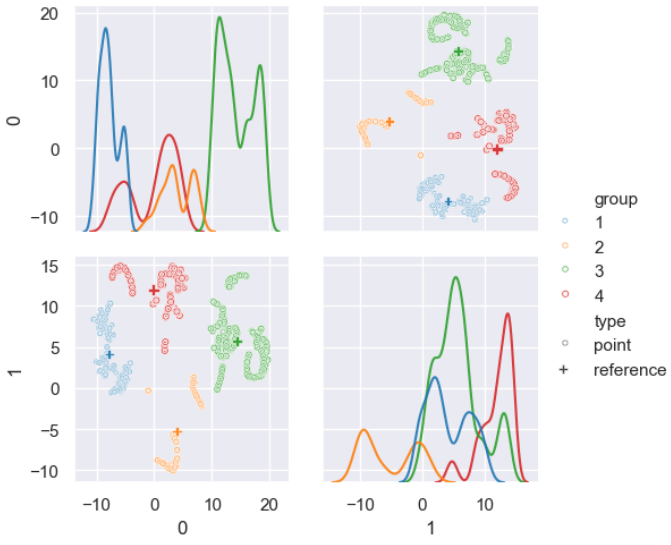


Fig. 3: 4 KNN Clusters on 2D UMAP Embeded Data

The final comparison focuses on the performance of the state-of-the-art HDBSCAN algorithm. Table V illustrates the respective indices resulting from the application of this clustering technique across the original dataset, MLP encoded dataset, and UMAP embedded dataset. Additionally, Figure 4 visualizes the classification generated from the original dataset, projected onto the UMAP embedded data, revealing an improved and refined data classification, despite not achieving the highest overall validation index.

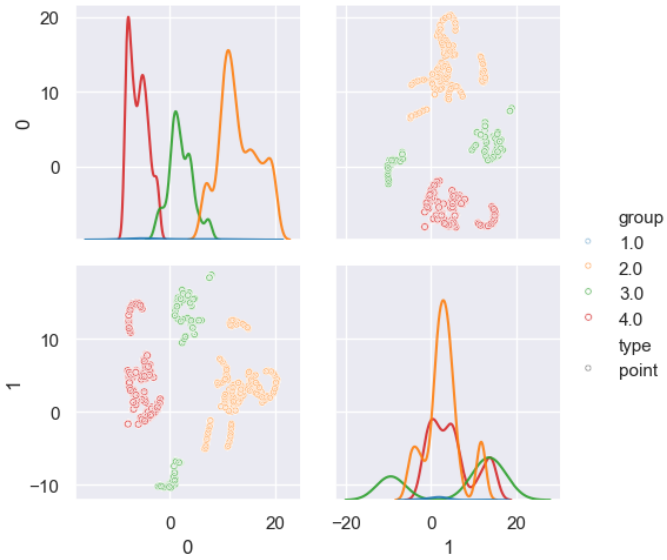


Fig. 4: 4 HBDSCAN Clusters on Original Data and UMAP embeded visualization

IV. Discussion and Conclusion

The limitations of naive algorithms in the classification process, which solely rely on data ordering, become apparent, particularly in scenarios where the subtleties and complexities of data patterns require more comprehensive analysis, as highlighted in the discussions on validation indices. Furthermore, the integration of fuzziness within the clustering process, as demonstrated by the Fuzzy C-NN algorithm, represents a significant advancement in accommodating data points with uncertain or ambiguous cluster assignments, thereby enhancing the adaptability and robustness of clustering analyses. Additionally, the Subtractive Density Cluster method emerges as a pragmatic and efficient choice, particularly suited for datasets characterized by well-defined clusters.

However, their efficacy may diminish when applied to datasets featuring diverse cluster shapes and densities, or substantial noise. The Billionaires dataset has proven to represent data not easily characterised and it is even possible, that the information provided is not enough to classify that people. These insights underscore the significance of employing advanced and adaptable clustering methodologies tailored to the specific complexities and nuances of the dataset under analysis. In this case, the large variety of results leads to conclude that further

metric	density algorithm	ra	K-Neighbors				Fuzzy C-Neighbors			
			k	davies bouldin	dunn	calinski-harabasz	k	davies bouldin	dunn	calinski-harabasz
coseno	mountain	0.6084	2	1.47E+14	0.00120	0.01457	2	3.13E+14	0.00118	0.01470
		0.0304	97	1.78E+13	0.00000	0.00019	31	7.19E+13	0.00245	0.10013
		0.0507	73	6.32E+12	0.00000	0.00010	31	1.81E+14	0.00387	0.20653
	substractive	0.3042	3	0.390	0.00012	0.00701	3	8.95E+13	0.00130	0.00488
mahalanobis	substractive	3.1171	4	0.929	0.00073	2.66542	4	2.442	0.00007	1.58882
manhattan	substractive	12.4119	2	0.680	0.08926	0.00474	2	0.605	0.00547	0.00062

TABLE II: Internal validation Indices of Best and Worst Clusters on Original Dataset

metric	density algorithm	ra	K-Neighbors				Fuzzy C-Neighbors			
			k	davies bouldin	dunn	calinski-harabasz	k	davies bouldin	dunn	calinski-harabasz
coseno	substractive	0.1455	4	205342.70	0.000085	0.003312	4	9.955E+14	0.000059	0.039574
		0.4367	2	0.149766	0.000083	0.009135	2	0.157344	0.000084	0.007941
		0.0218	198	8697.644	0	0.000029	91	3.77E+14	0.000152	0.210698
mahalanobis	substractive	4.3342	2	3.436317	0.058748	6.037185	2	2.152673	0.000051	4.426974
		2.6005	7	1.963734	0.000429	7.068504	7	3.636614	0.000165	2.833465

TABLE III: Internal validation Indices of Best and Worst Clusters on MLP Encoded Dataset

metric	density algorithm	ra	K-Neighbors				Fuzzy C-Neighbors			
			k	davies bouldin	dunn	calinski-harabasz	k	davies bouldin	dunn	calinski-harabasz
coseno	substractive	0.4164	3	1.63E+06	0	0.0099	3	6.43E+14	0	0.0131
		0.8328	2	0.288	0	0.0190	2	0.8857	0	0.0492
	mountain	0.2082	8	1.95E+04	0	0.0044	8	1.64E+14	0	0.1289
mahalanobis	substractive	0.7505	6	0.301	0.0010	0.0056	6	1.1783	0.00002	0.0015
manhattan	mountain	4.4826	4	0.624	0.0525	0.3972	4	1.3855	0.00370	0.8063
	substractive	4.4826	3	0.769	0.0103	0.5044	3	1.5911	0.00281	1.2767

TABLE IV: Internal validation Indices of Best and Worst Clusters on UMAP Embeded Dataset

Dataset	davies bouldin	dunn	calinski-harabasz	K
Original	0.65	0.00124	0.01256	4
UMAP	0.528	1.8E-05	0.053	28
Encoded	0.5156	1.40E-06	0.0036	22

TABLE V: Internal validation Indices of the HDBSCAN clustering on different Datasets

analysis is needed, and that unsupervised techniques alone are not enough to make hard conclusions on the dataset.

REFERENCES

[1] N. Elgiryi, "Billionaires Statistics Dataset (2023)." [Online]. Available: https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset

[2] O. L. Quintero M, "Machine Intelligence for Human Decision Making," *Editorial Artes y Letras*, Aug. 2019.

[3] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Feb. 2018. [Online]. Available: https://arxiv.org/abs/1802.03426v3

[4] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of the Intelligent and Fuzzy Systems*, vol. 2, pp. 267–278, Jan. 1994.

[5] V. Torra, "On the selection of m for Fuzzy c-Means." Atlantis Press, Jun. 2015, pp. 1571–1577, iSSN: 1951-6851. [Online]. Available: https://www.atlantis-press.com/proceedings/ifsa-eusflat-15/23735

[6] "M&MFCM: Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics," *Procedia Computer Science*, vol. 114, pp. 224–233, Jan. 2017, publisher: Elsevier. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917318689

[7] "Hierarchical Density-Based Spatial Clustering." [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.cluster.HDBSCAN.html

[8] R. Sirmen and B. Üstündağ, "Internal Validity Index for Fuzzy Clustering Based on Relative Uncertainty," *Computers, Materials & Continua*, vol. 72, no. 2, pp. 2909–2926, 2022, publisher: Tech Science Press. [Online]. Available: https://www.techscience.com/cmc/v72n2/47159

Fig. 5: Internal validation indices over differently parameterized clusters on the Original Data

Fig. 6: Internal validation indices over differently parameterized clusters of the encoded data

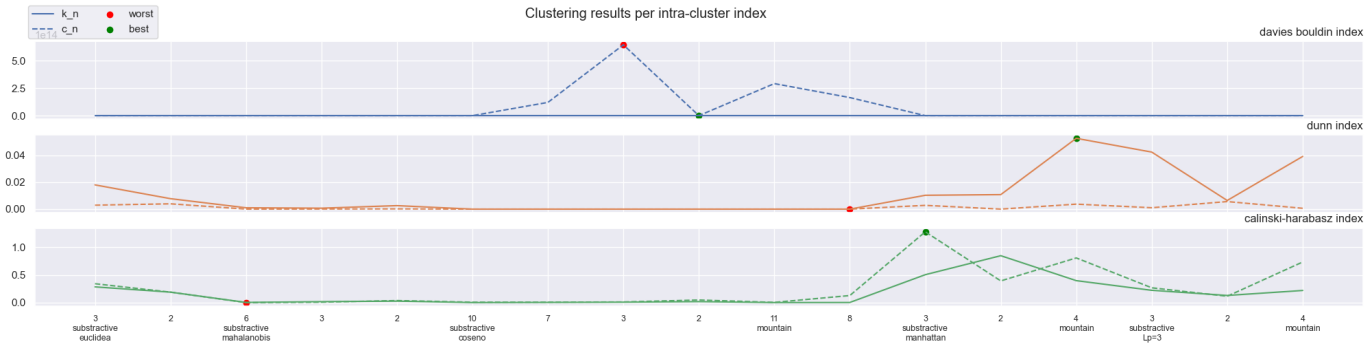


Fig. 7: Internal validation indices over differently parameterized clusters of the umap embeded data

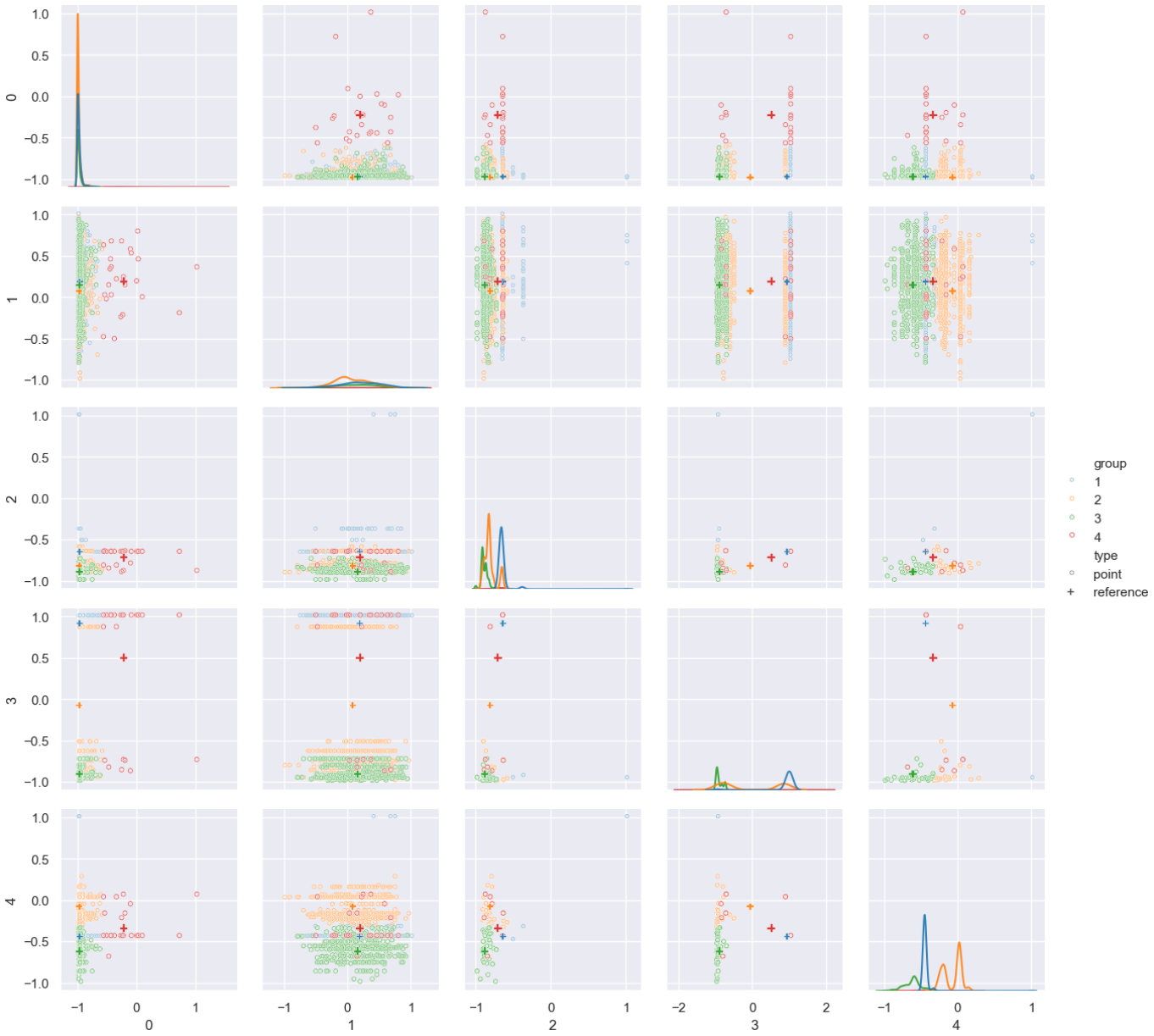


Fig. 8: 4 KNN Clusters on Original Dataset

Fig. 9: 7 KNN Clusters on 7D MLP Encoded Data