Robust Clustering Algorithm for High-Dimensional Real-World Data

Dr. Dávid Natingga
info@algohertz.com
AlgoHertz, Pražská 9056/9, 010 08, Žilina, Slovakia

August 12, 2024

Abstract

We present a robust high-dimensional clustering algorithm that outperforms clustering techniques available in popular machine learning libraries such as scikit-learn and TensorFlow. The algorithm evaluation and comparison was assessed using a rand index on the real-world disease datasets.

1 Challenge

Clustering high-dimensional real-world data involves significant challenges that complicate traditional approaches. These datasets are often noisy, corrupt, incomplete, and inaccurate, which can distort clustering results. The curse of dimensionality further exacerbates these issues, as data points in high-dimensional spaces become sparse and difficult to differentiate.

Additionally, real-world data frequently includes clusters of varying sizes and proportions across different dimensions, making it harder to accurately identify cluster boundaries. These challenges are especially prevalent in industries like finance, healthcare, telecommunications, utilities, and e-commerce, where high-dimensional data is the norm.

Existing algorithms often fall short when confronted with these complexities, leading to suboptimal clustering performance. Either the information is reduced through dimensionality reduction techniques, or the algorithms struggle to effectively utilize the inherent complexity of high-dimensional data.

2 Algorithm

Our algorithm belongs to the k-centroid clustering family and contains several key improvements:

- **Robust Semimetrics**: We utilize innovative semimetrics instead of traditional metric distances (Euclidean, Manhattan, Canberra), enhancing the algorithm's ability to handle complex and high-dimensional data structures.
- **Statistically Robust Centroid Calculation**: The algorithm preserves critical information while ensuring robustness in centroid calculations.
- **Smart Centroid Initialization**: Our method initializes centroids in a way that significantly improves clustering performance.
- **Annealing-based Centroid Convergence**: This technique avoids local optima and strives for global optima, ensuring more accurate clustering results.

The algorithm is designed to excel in real-world data scenarios, particularly in handling:

- High-dimensional clusters,
- Corrupt and incomplete data,
- Clusters with varying sizes and proportions across different axes.

Unlike many other algorithms, our solution operates directly on high-dimensional data without requiring dimensionality reduction techniques.

3 Evaluation Datasets

To evaluate the performance of the clustering algorithm, we used four publicly available microarray datasets by Borovecki [1], Gordon [2], Khan [3], and Sorlie [4].

In addition, we also used one *artificial* dataset that was created as follows. First, a random centroid for each cluster was generated in a high-dimensional cluster space. Afterwards, for each cluster and each dimension, a standard deviation was randomly generated. Afterwards, for each cluster, we generated normally-distributed random points.

The following table summarizes the datasets used for the algorithm evaluation.

Dataset	**Info**	**Size**	**Dimensions**	**Clusters**
Artificial	Normally distributed	898	500	5
Borovecki	Huntington's Disease	31	22283	2
Gordon	Lung Cancer	181	12533	2
Khan	SRBCT tumour	63	2308	4
Sorlie	Breast cancer	85	456	5

Table 1: Summary of Evaluation Datasets

4 Clustering Performance

We have conducted comprehensive evaluations comparing our algorithm against other publicly available clustering methods and various dimensionality reduction techniques from the popular machine learning libraries scikit-learn 1.5.1, TensorFlow 2.16.2, and hdbscan 0.8.37.

The dimensionality reduction techniques assessed include principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoders.

The clustering algorithms evaluated alongside our proprietary algorithm include k-means, k-medoids, k-medians, Gaussian mixture models (GMM), agglomerative clustering, spectral clustering, density-based clustering, and hierarchical density-based clustering.

The performance of our proprietary algorithm was assessed using the Rand Index, a measure of clustering accuracy. In the summary table below we can see that our algorithm performs better than any other publicly available clustering method assessed.

Dataset	**K-means score**	**Best public method**	**Best public method score**	**Our proprietary algorithm score**
Artificial	0.648	2D t-SNE k-medoids	0.851	0.999
Borovecki	0.548	3D PCA GMM	0.604	0.819
Gordon	0.522	2D PCA GMM	0.856	1.0
Khan	0.677	4D PCA k-medoids	0.731	0.748
Sorlie	0.801	Agglomerative clustering	0.867	0.905

Table 2: Clustering Performance Results

5 Availability

In order to use the algorithm, please, contact Dávid Natingga at info@algohertz.com.

References

[1] Flora Borovecki, Luca Lovrecic, Jingan Zhou, Hyung-Jun Jeong, Franziska Then, H Diana Rosas, Steven M Hersch, Peter Hogarth, Boris Bouzou, Robert V Jensen, et al. Genome-wide expression profiling of human blood reveals biomarkers for huntington's disease. *Proceedings of the National Academy of Sciences*, 102(31):11023–11028, 2005.

[2] Gavin J Gordon, Robert V Jensen, Louis L Hsiao, Steven R Gullans, John E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967, 2002.

[3] Jerald Khan, Javed S Wei, Markus Ringnér, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Curtis Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

[4] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Sverre Geisler, Hanne Johnsen, Trevor Hastie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.