Rediscovering Hierarchical Clustering

Jiahang Wu, Qihao Yun, Randy Zhang, Jimmy Yao

I. Introduction

Unsupervised machine learning is fascinating way to understand data sets without human interventions and bias. Hierarchical clustering were one of the earliest clustering algorithms developed in 1950s and 1960s. (Sibson, 1973) Hierarchical clustering is a particular case that elaborates the relative relationships of items with its nested clusters (a tree-like structure typically represented in dendrogram) which offers unique visualization values.

In modern discourse, hierarchical clustering algorithms seem to have gradually fallen out of favor due to its intensive computational demand. However, more and more modern researches took the concept of hierarchical clustering (using newer variants or self-defined methods) and achieved good results with a variety of data types. (Zhu and Guo, 2014)

In this project, we will explore the limits or capabilities of some classical hierarchical clustering methods as well as a modern, new method aiming to revitalize hierarchical clustering.

II. Overview

This project investigates a selection of hierarchical clustering algorithms by examining chiefly their 1) behavior or suitability with different types of datasets and 2) cost of deployment and complexity.

To evaluate the methods' interaction within different contexts, we have prepared a diverse pool of data sets which we categorize as labelled and unlabelled. The two groups of data sets will challenge methods' abilities in different data contexts. We will evaluate methods by internal and external (when applicable) clustering indices.

A. Literature

For this project, we have consulted a number of literature pieces both old and new. We accessed the works that have first proposed hierarchical clustering methods we are examining. Among them, Ward (1963) was perhaps the earliest paper that has clearly elaborated the concept, pseudocode, and applications of a general hierarchical clustering algorithms. In the same paper Ward also proposed his own implementation which is now known as the "Ward's" linkage that diverges from earlier methods in the use of function optimization. For our evaluation of the methods, a 2015 paper dedicated to comparing and testing hierarchical clustering methods offered some perspectives on discussing methods' performances. (Behaeghel, 2015) Finally, we read (though, only referencing three) a number of more recent researches using hierarchical clustering methods for our discussion section.

B. Methods

The general hierarchical clustering algorithm was first described clearly in J. H. Ward's "Hierarchical Grouping to Optimize an Objective Function" in 1963. Hierarchical clustering methods create nested clusters with either the agglomerative ("bottom-up") or the divisive ("top-down") approach by iterative merger or division.

The agglomerative approach assumes greatest amount of information is available by starting when all observations are ungrouped, start from individual items and iteratively merge clusters into larger ones. (Ward, 1963) Divisive methods start from a single cluster and iteratively divide it into smaller clusters. Generally, agglomerative methods excel in finding smaller clusters and divisive for larger clusters, and we will be focusing on this type of

We might consider the approaches in an analogy with an evolution tree which parallels the structure of nested clusters. In this analogy, the natural differentiation of species from a common ancestor may be regarded as the divisive process; while our modern reconstruction of this relationship by taxonomists is like the agglomerative process.

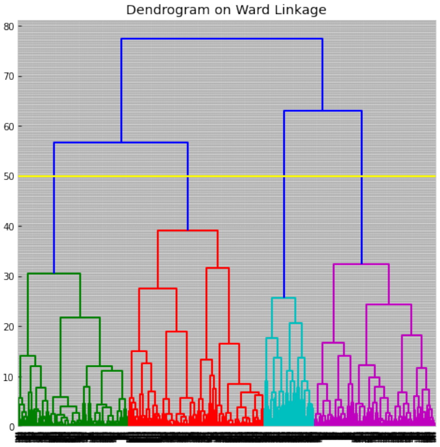


Fig. 1. A Dendrogram Example

The chief differences between the methods lie in their linkage criterions define how the method choose the next merger. Different distance metrics (Euclidean, Manhattan, etc.) can be applied for the calculation for some classical methods, but we will only use Euclidean as some methods we have chosen only accept Euclidean distance.

In this report, we will examine the following agglomerative methods:

SLINK (single linkage): An improved method to the nearest neighbor concept from the 1950s. SLINK performs

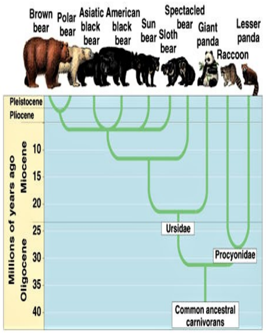


Fig. 2. An Evolution Tree

nearest neighbor (single linkage) merger . Its development was primarily aimed at optimizing the computation and storage processes of nested clusters and dendrogram, rather than to innovate the linkage. (Sibson, 1972)

Single linkage suffer from chaining effect as it may join clusters prematurely and becomes suspectable to initialization. (Sibson, 1972) The linkage is similarly sensitive to outliers and generally struggles on noisy or unclearly separated data sets. (Gagolewski, 2016) SLINK requires O(N) space and O(N2) computations, and it remains the least computationally demanding method among classical methods and a commonly used method currently for that reason.

Ward's Linkage: Commonly referred as the Error Sum of Square criterion, the Ward's linkage was developed in 1963 as an alternative to single linkage. Ward's method can be applied on hierarchical and non-hierarchical problems alike. The method can cluster regression models used in Air Force training without "serious loss" of accuracy while also performing well without optimizing for predetermined cluster numbers for non-hierarchical problems.

The linkage minimizes the change in the value of an objective function (typically ESS – sum of squared deviation about the mean) to minimize information loss. The complexity of this method depends on the implementation and function used.

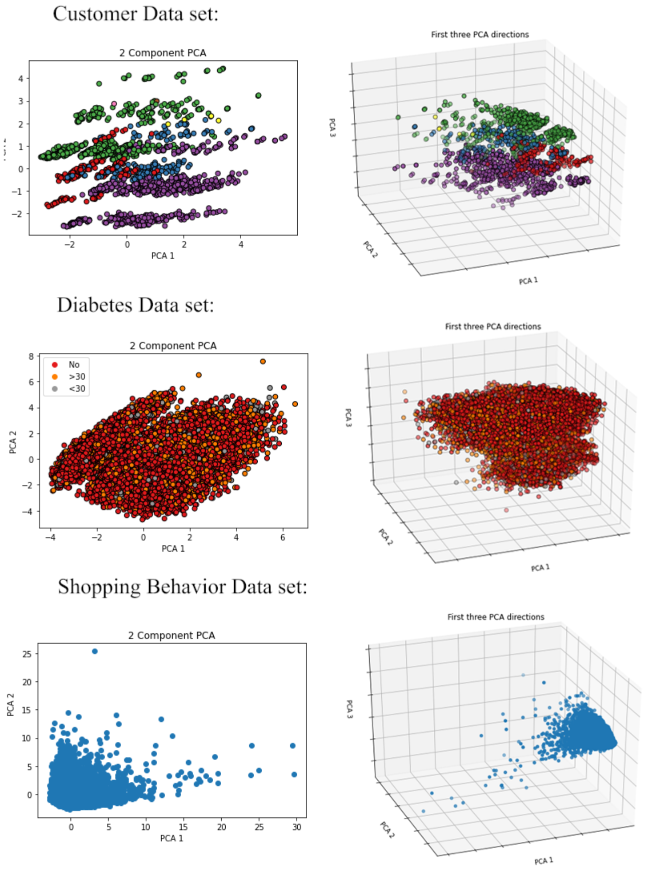
WPGMA (Weighted Pair Group Method with Arithmetic Mean) was first described in the 1958 paper by Sokal and Michener for a "statistical method" to evaluate taxonomical relations. (Sokal and Michener, 1958) The WPGMA algorithm constructs a dendrogram. The distance of a cluster i to another cluster j is simply the arithmetic mean of the average distances between members of i and j.

Centroid Linkage or UPGMC (Unweighted Pair-Groups Method Centroid) uses the distance between clusters' centroids as its criterion where a new centroid is computed after every cluster merge. This is very similar to the median linkage which differ primarily in the calculation of the new cluster's centroid.

GENIE is a new method developed in 2016 which proposed a new linkage based on economic inequality measures (e.g., Gini index) to evaluate the imbalance of cluster merges with a user-supplied threshold. Economic inequality indices fulfill the progressive transfer principle where transfers from concentrated (rich) to less concentrated (poor) does not incur growth in index value. (Gagolewski, 2016) The user supplied threshold can be important in some data sets as it determines how sensitive the method is to changes to overall equality in clusters. Low threshold can bar smaller clusters from being discovered; while, high threshold will result is similar behavior to SLINK method. The creator suggested (0.3,0.5) in general for gini-index. GENIE benefit from the well parallelized implementation of these indices and is able to scale well with different data types.

C. Data Sets

We separate the data sets used into two groups: labelled and un-labelled. These data sets are selected with the intent to challenge the methods' robustness against a varied pool of data sets. However, we are not including data sets with very mixed data types due to some methods' mandate for Euclidean distance function. The data sets are reduced here with PCA to demonstrate their distribution and shape.


Customer Data set:

Diabetes Data set:

Shopping Behavior Data set:

Fig. 3. 2D and 3D PCA of the Unlabelled Data sets

Unlabelled

Customer Dataset: demographic information of customers who entered a fast-moving consumer goods (FMCG) store. Textual, ordinal variables converted to numeric. Categorical variable one-hot into multiple features. Clean and contains no missing value. Shape: 2000 x 7. PCA shows some stratification with approx. 5-6 clusters. We expect these non-circular clusters to pose a particular challenge to some methods.

Diabetes Dataset: The dataset represents clinical care at 130 US hospitals and integrated delivery networks from 1998 to 2018, including over 50 features representing patient and hospital outcomes. Data was selected for encounters based on 5 criteria: inpatient, diabetic, length of stay at least 1 day and at most 14 days, lab tests were performed during the encounter, and medications were administered during the encounter. Data set is cleaned and preprocessed using label encoding. Selected numeric features directly related to diagnosis are categorized into levels. Shape: 71518 x 50. PCA shows 2-3 possible clusters. In order to make the dataset feasible to run, PCA with 19 components was applied. 19 components comes from how medication columns could be grouped together, and diag_1, diag_2, and diag_3, could also likely be grouped together.

Online Shopping Behavior Dataset: records of online store customers consisting of 18 columns related to shopping behaviors such as purchase frequency and current balance. Clean with no missing values. Shape: 8950 x 17. No clear cluster observable from PCA.

Labelled

Dry Beans Dataset: Dry beans with 16 numeric attributes describing the shape and form of the beans. 13600 x 16, 7 true labels.

MNIST Dataset: Images of handwritten digits (0-9), resized to 20*20 pixel (from 28*28) and represented as arrays. Final Shape 2000 x 400, contains 10 true labels.

Simple Artificial Dataset: a self-generated 2,000 row dataset composed of 1 uniformly and 3 normally distributed variables. Contains two well separated true cluster.

Complex Artificial Dataset: a self-generated 10,000 row dataset composed of 2 uniformly and 7 normally distributed variables. Contains three true cluster.

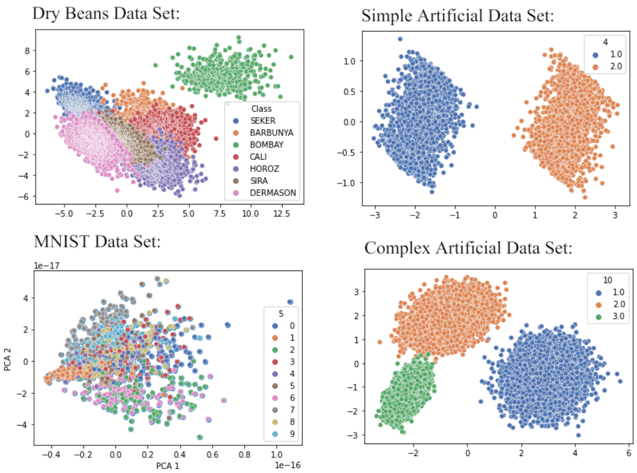
Note, the features' distributions in artificial data sets surround the predefined clusters. Hence, there should be no further data structure other than the clusters themselves.

While there are some packages available, we want to have more minute control over the generation of the artificial data. The data is generated by hand where for each feature, we define range, centrality for each cluster.

D. Evaluation

This study evaluates the clustering results of all methods with three internal indices: Silhouette, Calinski-Harabasz, and Davies-Bouldin. Where Calinski is the ratio of within and between clsuter variability; Davies-Bouldin is the ratio of in-cluster distances to centroid and inter-cluster centroid distances; silhouette is the confidence of the membership of items in a cluster. (Rendon et. al., 2011) Ultimately, these indices all aim to measure clean, well-separated clusters.

These indices primarily serve as references for our analysis. We intend to counter the shortcomings of indices with different data sets and methods by using an ensemble of three indices that are widely used and easily understood. As you will see, these indices – though all aiming to measure clean, well separated clusters – may give conflicting conclusions due to the complexity of the circumstances.


Dry Beans Data Set:

Simple Artificial Data Set:

MNIST Data Set:

Complex Artificial Data Set:

Fig. 4. 2D and 3D PCA of the Labelled Data sets

For unlabelled data sets in particular, there exist no well-recognized, universal method of evaluating the significance of the resulting clusters. (Kimes et al., 2017) Procedures are created and investigated for individual methods, but because of the difference in linkage criterion, procedures do not necessarily extrapolate for other methods. These procedures are often too complex for the scope of this report to be addressed fully.

We will compare the results of the methods. Agreement and disagreements between them will contribute to our understanding of the methods and hierarchical clustering's applicability as a whole. We also hope that we may extend the results – if some agreement or pattern emerge – to further explore the data sets and validate whether underlying data structures are revealed by the methods.

For labelled data sets, we have a mix of real world and artificial data sets that are relatively unambiguous (having no hidden structures otherwise). For these data sets, they will serve to compare the robustness of the methods by comparing the generated clustering with priori (true labels).

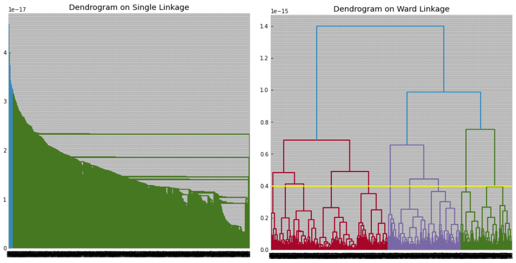


Fig. 5. Dysfunctional SLINK (left) and Ward (right) on MNIST

For the sake of brevity, we will not include dendrogram and internal metric score outputs selectively. In our context of appplication, dendrogram's role lies primarily in helping us determining whether the method is performing adequately and the appropriate number of clusters. A poor dendrogram indicates the method's inability to perform on the data set. See the following two images for comparison.

The dysfunctional dendrogram shows the method struggles to really tell clusters apart, resulting in large, early clusters that simply absorb the remaining clusters.

For the unlabelled data sets, we are chiefly concerned with determining the "optimal" cluster number which we can determine by combining the dendrogram and the scores. The results on the right are the outputs for the customer dataset. Sample results are shown in the order of SLINK-Ward-WPGMA-Centroid-GENIE, from top to bottom.

On the left column is the dendrogram; on the right, the scaled internal metrics' scores for the number of clusters shown in the dendrogram. For Silhouette and Calinski, we take the highest point as the optimal number, while Davies-Bouldin uses the lowest score as optimal.

These results are tabularized for each individual datasets which we will disucss in the results section. N/A in these tables indicate dysfunctional results from either the plot or the metric.

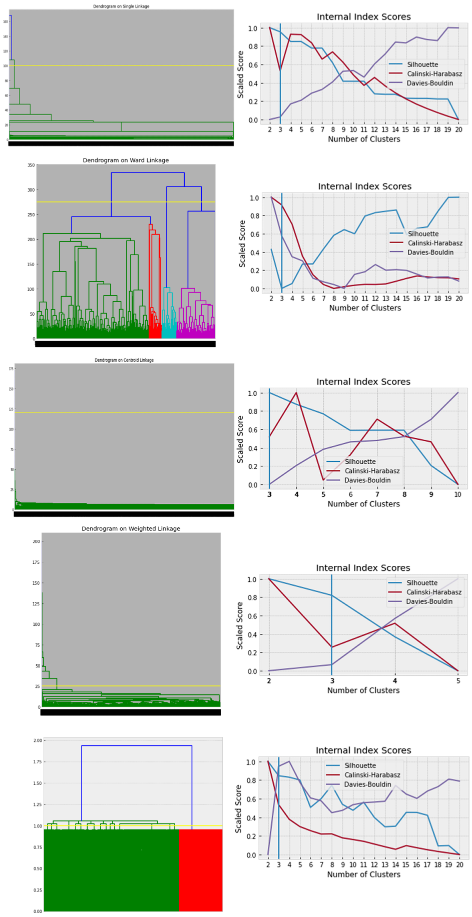


Fig. 6. Sample Dendrogram Outputs (Diabetes Data Set)

III. Results

A. Customer Dataset

TABLE I

Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
49.2	48.2	47.9	48.2	0.430

TABLE II

Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	8	3
Ward	20	4	18
WPGMA	147	177	177
Centroid	143	173	173
GENIE	29	2	15

The results for WPGMA and Centroid show that a greater number of clusters (greater than 100) would be ideal. However, these large numbers do not necessarily point to the fact that over 100 clusters would be the ideal number of clusters. There were only 2000 customers to cluster, meaning that at least one cluster would contain less than 5% of the 2000 clusters. It is likely that the methods were dysfunctional and did not actually find proper clusters.

Otherwise, the optimal number of clusters from the internal metrics between SLINK, Ward, and GENIE do not appear to be inclined towards any number of clusters: some suggest less than 5, others metrics suggest more than 15. In addition, it can be seen that the internal metrics seem to conflict with each other. For Ward and GENIE, Silhouette score and Davies-Bouldin index return greater number of clusters, while Calinski-Harabasz score returned a lower number of clusters. For the Ward method, exactly the opposite occurred.

Examining the dendrograms for SLINK, Ward, and GENIE methods, we can see that for SLINK, the optimal threshold appears to either be 1.9 or 2.3 with 2 or 9 clusters respectively. For Ward, the optimal threshold appears to be 50 for 4 optimal clusters. For GENIE, the optimal threshold appears to be 1 for 28 clusters. These values to match some of the values we obtained through the internal metrics, but still conflicts with each other.

With how varied much of the data is, we can only conclude that the number of clusters that each method suggests, and each internal metric, are varied.

In terms of time cost, GENIE is much faster than the other methods when applied to the customer data set with a time cost of less than 1 second, while the other methods are equally as slow with time costs of approximately 48 seconds each.

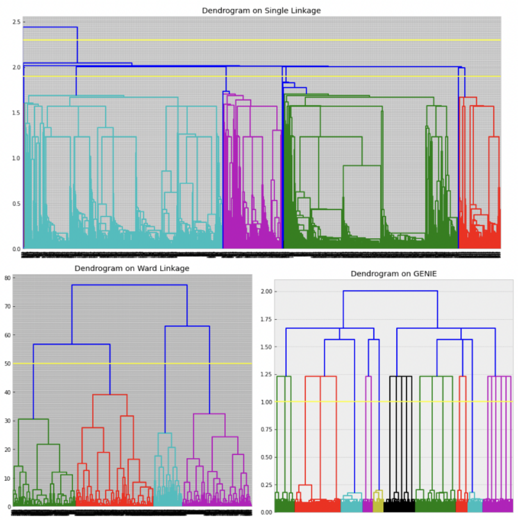


Fig. 7. Slink Dendrogram (top), Ward Dendrogram (bottom left), GENIE Dendrogram (bottom right)

B. Diabetes Dataset

TABLE III
Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
958	1052	1121	948	14.4

TABLE IV
Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	2	2
Ward	19	2	9
WPGMA	2	5	2
Centroid	2	4	2
GENIE	2	2	2

The results of the ensemble internal metrics for each method when applied to the Diabetes data set appear to point to 2 clusters being the optimal number of clusters. SLINK and GENIE methods, in particular, have the same value of 2 throughout the 3 evaluation metrics. However, for the Ward method, the Silhouette and Davies-Bouldin indices suggest the number of clusters should be greater than 5, greater than what the majority of the other indices suggests, showing that Ward, had perhaps performed poorly on this data set.

In terms of time cost, GENIE is much faster than the other methods when applied to the Diabetes data set, with a time cost of less than 15 seconds, while the other methods are equally as slow with time costs of approximately 1000 seconds each.

C. Behavior Dataset

TABLE V
Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
53.7	83	34.1	34.8	1.4

TABLE VI
Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	3	2
Ward	2	2	9
WPGMA	>20	19	20
Centroid	>16	6	12/14
GENIE	2	2	2

The results of the shopping behavior dataset turn out to be more complicated. By looking at the time cost of each method in second table, Ward took the longest to run (83s), whereas GENIE only took 1.4 seconds. In particular, only GENIE appeared to be robust that the optimal clusters have the same value of 2 throughout three evaluation indexes. SLINK and Ward also show the result that the optimal cluster number should be 2. However, WPGMA and Centroid return different numbers of optimal clusters for all three evaluation indexes and none of these values are close to 2. By looking at the Ward, we can know that optimal clusters = 2 is more realistic. Therefore, we can conclude Ward and GENIE has relatively better performance.

D. Dry Beans Dataset

TABLE VII
Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
344	349	449	340	3.16

TABLE VIII
Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	2	2
Ward	3	2	3/4
WPGMA	3	4	3
Centroid	2	3	3
GENIE (t=0.2)	2	2	3

SLINK and WPGMA reuslts are visibly dysfunctional. Davies-Bouldin index returns no results for the data set. Centroid clustering result appears to have clear clusters, but upon further inspection, similarly contains a large cluster absorbing very small ones.

Ward's dendrogram suggests a clear inclination towards two clusters as the nested clusters are evenly broken down in small groups. GENIE appears more promising, having identified several clearly separated clusters that are resistant to further changes (tall bars).

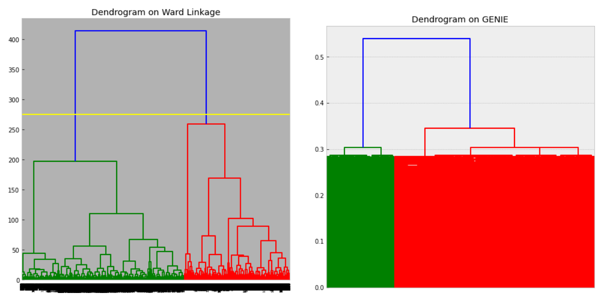


Fig. 8. Ward (left) and GENIE (right) Dendrograms

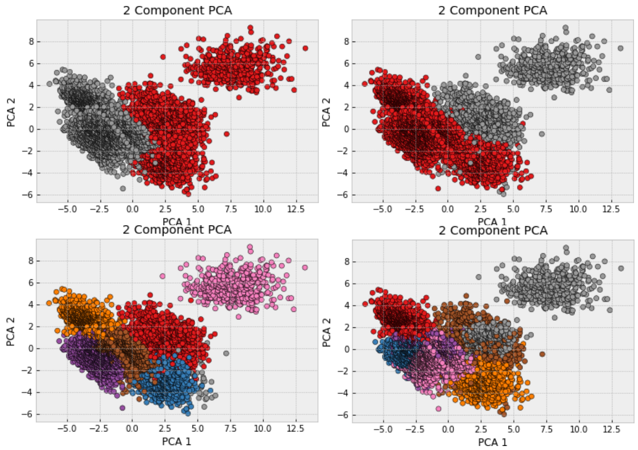


Fig. 9. Comparison of Ward (left) and GENIE (right) Clustering; n=2 (top) and n=7 (bottom)

At the optimal cluster numbers suggested by internal indices (n=2), both methods shows clear understanding of real class boundaries. However, when attempting to find n=7 clusters, the results, while observing some class boundaries, are too messy to be meaningfully applied with external indices. We did not find existing taxonomical literature about the beans as they appear to be variants of the common bean (Phaseolus vulgaris) cultivated primarily in Turkey.

E. MNIST Dataset

TABLE IX

Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
48.8	48.4	46.2	44.9	0.54

TABLE X

Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	2	N/A
Ward	3	2	N/A
WPGMA	5	>20	N/A
Centroid	2	2	N/A
GENIE (t=0.2)	2	2	N/A

SLINK and centroid reuslts are dysfunctional. Davies-Bouldin index returns no results for the data set. WPGMA clustering results shows some clear clusters, but the internal indices are contradictory.

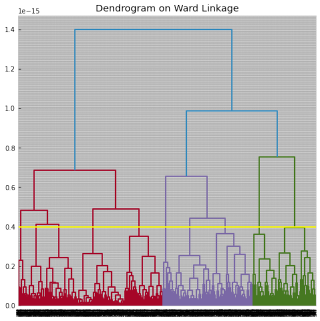


Fig. 10. Ward Dendrogram

At cluster = 2, GENIE results show a group of mostly 4 and 9 and another of the other digits. At higher cluster number, the pattern is progressively weaker. Ward shows no visible pattern at 2 clusters. At cluster = 10, clusters from both methods behaves similarly. Each cluster contains primarily one or two digits, but overall, they are clearly no where near a predicative model.

F. Simple Artificial

TABLE XI

Time Cost of Each Method in Second

SLINK	Ward	Weighted	Centroid	GENIE
149	150	149	149	0.33

TABLE XII

Optimal Cluster Number Per Internal Metrics

Method	Silhouette	C-H	D-B
SLINK	2	2	2
Ward	2	2	2
WPGMA	2	2	2
Centroid	2	2	2
GENIE	2	2	2

All methods performed with 100% accuracy on this data set. Simple Artificial dataset mainly works to detect clustering strategies for different methods. All methods result in a clear two-cluster model, as intended by the Davies-Bouldin index, Silhouette index, and Calinski-Harabasz index. SLINK functions as a cluster-adding method, which clusters the dataset by adding small single clusters to broad division. While, WARD functions by dividing clusters hierarchically. Intending by the dendrogram, the result of SLINK looks like a sliding hill while WARD looks like a probability tree diagram. GENIE implied a similar strategy as WARD but costs 500 times less time. Centroid and WPGMA's dendrograms indicate that both strategies are used since the cluster adding and dividing are detected at the same time.

G. *Complex Artificial*

TABLE XIII
TIME COST OF EACH METHOD IN SECOND

SLINK	Ward	Weighted	Centroid	GENIE
731	763	792	821	7.3

TABLE XIV
OPTIMAL CLUSTER NUMBER PER INTERNAL METRICS

Method	Silhouette	C-H	D-B
SLINK	2	2	>20
Ward	3	3	3
WPGMA	2	2	2
Centroid	2	2	5
GENIE	3	3	3

All methods have discovered 3 clear clusters at some stage of the computation. Ward and GENIE remain relatively robust that the optimal clusters are the same as the real cluster number (n=3). SLINK and Centroid methods, though arriving at three clusters, behave

Ward and GENIE have 100% accuracy. SLINK, Centroid, and WPGMA only see 2 clusters. Their resulting labels when specified n=3 do not recognize a distinct third cluster, such as the WPGMA results when n=3: (1.0 10000), (2.0 6246), (2.0 3754 3.0 10000).

H. *Summary*

Ward and GENIE perform more consistently across various data set. A key observation is that the linkage of ward and genie are analogous in that they both consider clusters against some function - Ward ESS, and GENIE Gini index - evaluate clusters within the context of the larger data set. In our results, the two methods behaves similarly, often recognizing distinct, medium sized clusters early on.

SLINK's behavior and results are similar to Nearest Neighbor. The general behavior of centroid and WPGMA are similar: producing many lower level, small clusters that are subsequently merged into larger clusters. Interestingly, the two methods, when merging smaller clusters, often choose clusters that appear distant in dendrograms, forming a web-like appearance that is hard to read. These two methods, along with SLINK, often struggle to recognize lower level data structure whenever the data set becomes less clearly-separated in PCA view. They tend create several large clusters that progressively absorbs tiny (<10 observations) clusters. Consequently, they do not appear to be recognizing the relative relationships of the data and often produce diagrams that are hard to read.

The internal indices turned out to be poor indicators for our study. They almost always converge around 2-4 clusters for any data sets, which appears irrelevant as we intend to see if the methods can recognize less obvious structures within the data sets. While we have initially planned for applying external indices, they often cannot be meaningfully applied to real world data sets as the resulting clusters at the true cluster number usually do not separate according to true labels. This may be a consequence of low correlation between the true label and some variables of the data set.

In a very recent study, traditional hierarchical clustering methods did have some successes; however, the study examined relatively small data sets and used pearson correltaion as the similarity (distance) metric. (Xu et. al, 2022) Our study of the methods, for the sake of consistency, has chosen to use Euclidean metrics only, and that may be the at the heart of the issue for us.

IV. DISCUSSION

Computational complexity is not the primary barrier to classical hierarchical clustering methods; **the lack of understanding of their theoretical basis and their limitations is.** This research started with an optimistic expectation of clustering methods as a somewhat magical technique, but the blind applications of techniques are often unrewarding.

Even in very recent studies, we sometimes still see groups applying algorithms (namely, hierarchical clustering methods) without fully contextualizing or understanding the mechanism of the technique. One study did not even bothered to note what linkage was used for their dataset. (Ebrahimi, 2022)

It can be unrealistic to expect clustering algorithms to behave with the precision of a predicative model or to expect to magically reveal "hidden" data structures. In the dry beans data set, the many distinctions of beans were not measured or possibly included in the data set; perhaps there are nutritional distinctions between beans; or perhaps the beans are indeed fundamentally similar. So the human perceived differences often are not communicative to the methods.

Hierarchical clustering methods do have the advantage to show the reasoning (relationship) which we humans can have a hard time communicating. Understanding the limits and capabilities of the algorithms are essential to using the methods. Expectations and hypotheses need to be based on the context of the methods as well as the data sets.

From the experiments, we also confirmed the limitations of internal indices and the necessity to use an ensemble of them for any clustering application. Because they ultimately value clean, well-separated clusters, their applications in comparing clusters and true clusters can be very limited as a result of chaos in real classifications. We also discovered that some indices (e.g. silhouette) can be computationally very expensive, potentially more so than the hierarchical clustering themselves.

One notable remark about internal indices is that they often do not reflect poorly performing results. As discussed, poor clustering results typically have a large, early cluster the progressively absorbs much smaller clusters. On those cases, we observe good score - often unanimously - on 2 or 3 as the "optimal" cluster number. Again, this shows the severe limitations of internal indices, and a selection of optimal cluster number would require careful human inspection of the actual results and data sets.

V. Division of Labor

TABLE XV

Time Cost of Each Method in Second

Jimmy Yao	Implementation and standardization of coding workflow. Analysis of Customer, Diabetes data sets
Jiahang Wu	Analysis of Dry Beans, MNIST, Complex artificial data sets. Structuring and writing of the report. Reading and incorporation of literature pieces Generation of artificial data sets.
Horace Yun	Analysis of Simple Artificial Data set Identification of literature pieces
Randy Zhang	Analysis of Shopping Behavior data set. General information of the five clustering methods.
All	Writing and presenting in-class presentation Identification of 1-2 real data sets

References

[1] Behaeghel, Isabelle, Anouk Veldhuis, Libo Ren, Estelle Méroc, Frank Koenen, Pierre Kerkhofs, Yves Van der Stede, Jacques Barnouin, and Marc Dispas. “Evaluation of a Hierarchical Ascendant Clustering Process Implemented in a Veterinary Syndromic Surveillance System.” Preventive veterinary medicine 120, no. 2 (2015): 141–151.

[2] Ebrahimi, Pooria, Annalise Guarino, Vincenzo Allocca, Stefano Caliro, Rosario Avino, Emanuela Bagnato, Francesco Capecchiacci, et al. 2022. Hierarchical clustering and compositional data analysis for interpreting groundwater hydrogeochemistry: The application to campi flegrei volcanic aquifer (south italy). Journal of Geochemical Exploration 233 (February 2022): 106922.

[3] Kimes, Patrick K, Yufeng Liu, David Neil Hayes, and James Stephen Marron. “Statistical Significance for Hierarchical Clustering.” Biometrics 73, no. 3 (2017): 811–821.

[4] M. Gagolewski, M. Bartoszuk, and A. Cena. Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. Information Sciences, 363:8–23, 2016. doi:10.1016/j.ins.2016.05.003

[5] M. Gagolewski. genieclust: Fast and robust hierarchical clustering. SoftwareX, 15:100722, 2021. doi:10.1016/j.softx.2021.100722

[6] R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method" . The Computer Journal. British Computer Society. 16 (1): 30–34. doi:10.1093/comjnl/16.1.30.

[7] Rendón, Eréndira, et al. "A comparison of internal and external cluster validation indexes." Proceedings of the 2011 American Conference, San Francisco, CA, USA. Vol. 29. 2011.

[8] Sokal, Michener (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin. 38: 1409–1438

[9] Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236–244.

[10] Xu, Na, Chuanpeng Xu, Robert B Finkelman, Mark A Engle, Qing Li, Mengmeng Peng, Lizhi He, Bin Huang, and Yuchen Yang. “Coal Elemental (compositional) Data Analysis with Hierarchical Clustering Algorithms.” International journal of coal geology 249 (2022): 103892–.

[11] Zhu, Xi, and Diansheng Guo. 2014. “Mapping Large Spatial Flow Data with Hierarchical Clustering.” Transactions in GIS 18 (3): 421–35. doi:10.1111/tgis.12100.