International Journal of Engineering & Technology, 9 (4) (2020) 887-899

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data analysis using representation theory and clustering algorithms

Dr. Suboh Alkhushayni [1] *, Taeyoung Choi [1], Dr. Du'a Alzaleq [1]

[1] Computer Information Science Department, Minnesota State University, Mankato, Mankato, MN, 56001, USA
**Corresponding author E-mail: Suboh.alkhushayni@mnsu.edu*

Abstract

This work aims to expand the knowledge of the area of data analysis through persistence homology and representations of directed graphs. To be specific, we looked for how we can analyze homology cluster groups using agglomerative Hierarchical Clustering algorithms and methods. Additionally, the Wine data, which is offered in R studio, was analyzed using various cluster algorithms such as Hierarchical Clustering, K-Means Clustering, and PAM Clustering. The goal of the analysis was to find out which cluster's method is proper for a given numerical dataset. We tried to find the agglomerative hierarchical clustering method by testing the data that will be the optimal clustering algorithm among these three; K-Means, PAM, and Random Forest methods.

By comparing each model's accuracy value with cultivar coefficients, we concluded that K-Means methods are the most helpful when working with numerical variables. On the other hand, PAM clustering and Gower with Random Forest are the most beneficial approaches when using categorical variables. These tests can determine the optimal number of clustering groups, given the data set, and by doing the proper analysis. Using those the project, we can apply our method to several industrial areas such that clinical, business, and others. For example, people can make different groups based on each patient who has a common disease, required therapy, and other things in the clinical society. Additionally, people can expect to get several clustered groups based on the marginal profit, marginal cost, or other economic indicators for the business area.

Keywords: *Representation Theory; Data Analysis; Persistence Homology; Agglomerative Hierarchical Clustering; K-Means; Cosine Distance; Manhattan Distance; Minkowski Distance; Single Cluster; Complete Cluster; Average Cluster.*

1. Introduction

As society continues to become more technologically advanced, data collection has become significantly easier and is done in almost every facet of life. We can collect data on nearly anything, from our favourite sports team's performance to the propagation of specific strains of the flu. In most instances, data collection isn't as much of a barrier as knowing how to interpret that data and find what is relevant in each data set. This is where data science and analysis come into the picture. Many researchers are working on varied techniques that will help to analyse massive data sets and squeeze out whatever relevant data they can muster. Fundamental statistical analysis has been done for years. Still, in many cases, the tests that have been around for decades can't keep up with the sheer volume and magnitude of the data sets we have available.

One approach used for much larger data sets is to look for clusters within the data. We search for occurrences that are similar to one another and look for large sets of data points that can be grouped. However, with this method, many questions need to be answered. What scale should be used when searching for clusters? How close should the points be together to be considered related? What happens if there is a large cluster of data with a gaping hole right in the middle of it?

As an example sourced from Figure 1, consider the collection of symbols found in figure 1 of appendix A. What would your response be if someone asked you, "What information can be gathered from that data set?" There are three natural answers; a single letter A, eleven B's, or 176 points. But which one of these answers, if any, is relevant to the person who asked the question?
In different data sets, the same thing can happen. If you are looking at a small scale, the clusters appearing are different than those performing at a larger size. It is difficult to know what extent to use to get any data, and even more challenging to decide which scale gives relevant data. This is where a field of study within mathematics called persistence homology comes into play.

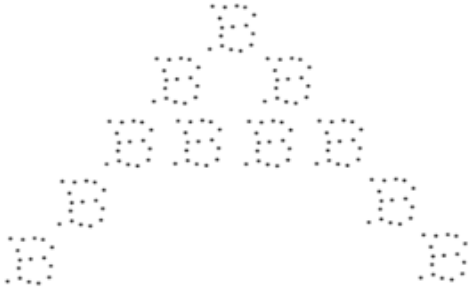

Fig. 1: Sample Data Set.

Without diving into too much jargon and detail, persistence homology allows us to look at all scales at once. We form a continuum of increasing sizes, and at each scale, we look for the clusters that exist. Additionally, as we "zoom out," some groups will combine with others to form larger clusters of data, like how the 176 points in our example cluster into 11 B's and eventually into a single large A. Persistence Homology shows us which groups "persist" as they absorb other clusters, and it gives a good idea of what scales cause a change in the clusters in the data [1].

Another thing is that persistence homology does exceptionally well is handling holes in the clusters within the data set. For example, in the 11 B's in figure 1, the two holes that help to make up the letters are large gaps with no data points in them. Some approaches used for finding clusters would have issues with these holes and may struggle to cluster the points in each B together. Another strength of these methods is that they are quite stable under small perturbations in the data set. If the points move slightly, the information that can be garnered from the Persistence Homology will stay the same, allowing for some variance in the data set to occur without skewing the results.

2. Methodology

When studying clusters of data using persistence homology, we look at varying scales and view how clusters of data combine into larger clusters or vanish as we increase our scale. This information can be encoded into what is referred to as a directed graph. A directed graph is just a collection of points connected by arrows. In persistence homology, the directed graph initially used is a very well behaved one that consists of dots all equally spaced apart and lying on a single horizontal line. The only arrows are the ones that connect adjacent dots from left to right.

An example of this can be seen in figure 2. To connect this to the idea of varying scales, imagine that as we move from dot to dot, from left to right, our scale of the data set is increasing, and the arrows are assigned maps that tell us how the clusters merge and disappear as we move up in size.

Fig. 2: Directed Graph Traditionally Used for Persistence Homology.

The encoding of data in this manner creates a mathematical object called a representation of a directed graph. From there, we break this object into the indecomposable representations, or smallest pieces. In this case, where the graph is just a line, as in figure 2, a finite number of these are the most minor parts that we need to work.

We analyze a data set by forming several clustered groups, noticing that the degree of error might increase or decrease. In Figure 3, we can see the determination of the clustering groups based on the size of epsilon (ϵ) which represents the error parameter [2].

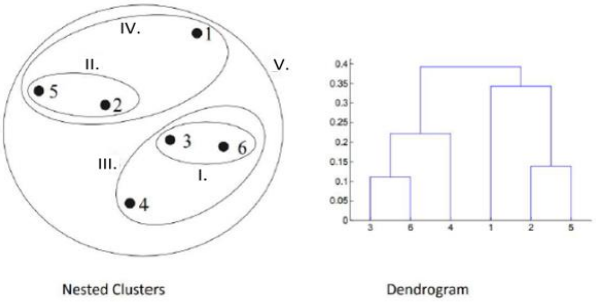

Fig. 3: Method of Clustering Using Persistence Homology.

Each point through 1 to 5 represents a group of data set. Points 3, and 6 are combined as one cluster (cluster I) because the epsilon parameter (ϵ) equals to 0.1. Points 2 and 5 also forms another cluster (cluster II) with has an epsilon that equals to 0.13. We can then see that the cluster I merged with point 4 as one big cluster (cluster III) with an epsilon value of 0.22. Additionally, cluster II merged to point 1 as another big cluster (cluster IV) with epsilon value of 0.33. Finally, the two clusters III and IV are combined as the most prominent cluster V with the epsilon value of 0.4.

Using the above Hierarchical Clustering method, we can get an insight into how to analyze different clustered groups based on the degree of epsilon values. For using this algorithm, we need to consider the distance of several data points to determine the value of the error parameter (ϵ). There are three different types of methods; single cluster, complete cluster, and average cluster that helps us performing that. In the Single Hierarchical Clustering, the distance between two clusters is the shortest distance between two different points in each cluster. In Figure 3, the shortest distance is between points 2 and 3 [3].

In a Complete Cluster, the distance between two clusters defined as the longest distance between two random points among the clustered group. For example, we can see two different complete clusters, points 1 and 4, and points 5 and 6.

Last, is the Average Cluster, which defines the distance between two clusters as the average distance between each point in one cluster to every point in another cluster. For example, the distance between clusters III and IV is calculated as the average of distances from points 1, 2, and 5 to the points 3, 4, and 6.

We then focused on the agglomerative method, which builds clusters from the smallest to the largest. After that, we obtained the value of the error parameter from the biggest to the smallest cluster to analyze the data set using the method of K-means. It is one of the simplest and popular unsupervised machine learning algorithms [7]. First, we randomly initialize k points, called means. Second, we classified each item to its closest mean, and we updated the mean's coordinates, which are equal to the average of the distances between the points that were categorized in the current mean. Last, we repeat the process for a given number of iterations, and at the end, we got our desired cluster. In general, the K-means approach is performing faster and more precisely than Hierarchical Clustering if K's values are large enough. The K-means approach also produces tighter clusters than the Hierarchical Clustering approach, especially if the clusters are enormous [7]. However, it is difficult to predict the K value. Moreover, if there is a global cluster, it does not work well [7].

Based on the above pros and cons of the K-means approach, we decided to use the Agglomerative Hierarchical Clustering algorithm to analyze data and develop how we can get insight from this method [9]. As in Figure 4, we notice that as the degree of single cluster incremented (from top to bottom), the individual data sets clustered as one (Letter A shape).

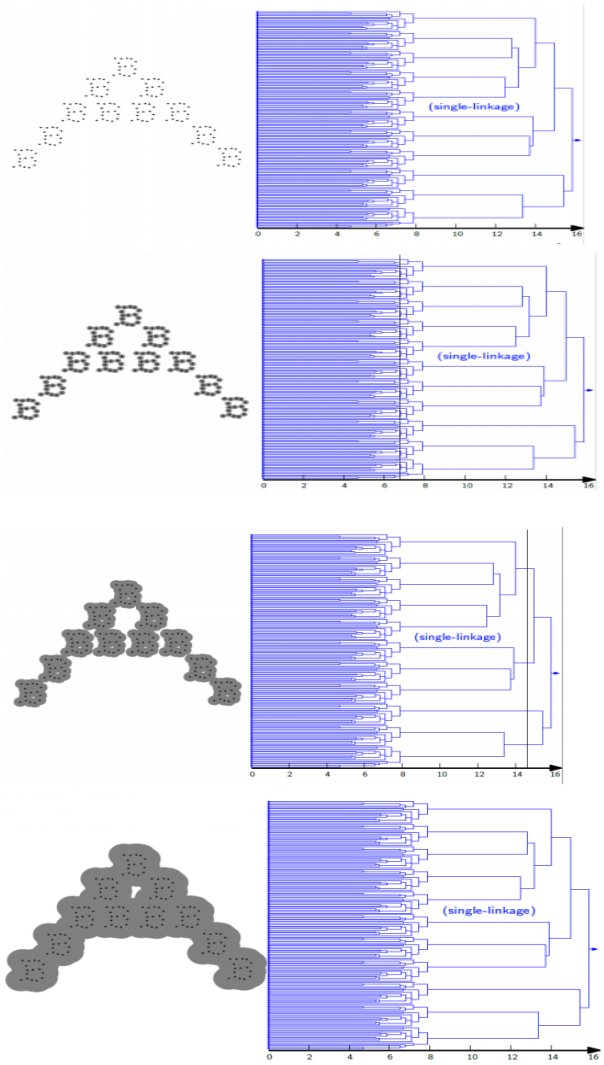

Fig. 4: Directed Graph Traditionally Used for Persistence Homology Epsilon =0, 6.2, 14.3, 16 from Top to Bottom

2.1. Hierarchical clustering

The Hierarchical Clustering method builds a grouping data set based on the dissimilarity measure concept. There are five different types of Hierarchical Clustering methods as described in Table 1 below.

Table 1: Methods of Hierarchical Clustering [17]

Method	Description
Ward Clustering	Minimize their sum of squared error with total within-cluster variance
Complete Clustering	Maximize the distance between one of the groups with a point and the other group with another point
Single Clustering	Minimize the distance between one of the groups with a point and the other group with another point
Average Clustering	Average the distance between one of the groups with a point and the other group with another point
Centroid Clustering	The distance between two points which are in each cluster

2.2. K-means clustering

To build a K-means clustering, we need to define the number of clustering groups first. Then, the algorithm will run until one data set is grouped in K number of clusters. The optimal results of this algorithm minimize the sum of squares of Euclidian variance. The steps of K-Means Clustering as the below [14],

1) Define the number of K clusters,
2) Reset the amount of K clusters with the starting average,
3) Repeat the above steps,
   a) Build a number of K clusters using each of the data set to make them closer.
   b) Each of the middle points is the new average
   c) Works until each of the central points of the K clusters will not be changed anymore.

2.3. PAM (partition around medoids) clustering with Gower dissimilarity coefficient

This method is specialized in the complicated data set, which contains all the unstructured data variables properly and well-organized. To be specific, the data set is in a nominal form, ordered, partitioned, and fractional. This method is similar to the K-means Clustering approach, but we first need to clarify its pros and cons. The advantages of using this method are,

1) It was easily inputting the complex data set using the dissimilarity matrix.
2) The Euclidean distance method uses the summation of dissimilarity to find more specialty and dissimilarity coefficient [18].

Gower Dissimilarity Coefficient compares each of the clusters' contributions to calculate the dissimilarity coefficient. If I and J are data sets, then their Gower Dissimilarity Coefficient is $S_{ij} = \frac{\mathrm{sum}(W_{ijk} * S_{ijk})}{\mathrm{Sum}(W_{ijk})}$.

For this case, S_{ijk} represents the number of k contributions. W_{ijk} takes the number of the binary value of 1 if the number of k is valid; otherwise, it is a binary of 0. Regarding the ordered and consecutive variables, when r_k has the k^{th} value in a given interval, then $S_{ijk} = 1 - \frac{|x_{ij}| - x_{ij}}{r_k}$ can be defined. If the data set is nominal, then $x_{ij} = x_{jk}$, and $S_{ijk} = 1$, otherwise it is 0. If the data set is binomial, then S_{ijk} represents + and – regardless each property exists or not. We summarized the above information in Table 2.

Table 2: Value of the Attribute K [6]

Variables	Value of Attribute K			
Case i	+	+	-	-
Case j	+	-	+	-
S_{ijk}	1	0	0	0
W_{ijk}	1	1	1	0

3. Case study and results

Italian Sommelier data will be analyzed using the above clustering method. The test of becoming Sommelier is hard and has a high failure rate. One person would like to be a Sommelier and investigate the knowledge of Italian's Wine structure using the 'wine' data set with R studio.

We used the following library packages from R studio: ‘cluster’, ‘compareGroups’, ‘HDclassif’, ‘NbClust’, and ‘sparcl’.

```
Library (cluster)
Library (compareGroups)
Library (HDclassif)
Library (NbClust)
Library (sparcl)
```

Then, we make a 'build a wine' data set from the above library.

```
'data.frame':   178 obs. of  14 variables:
 $ class: int  1 1 1 1 1 1 1 1 1 1 ...
 $ V1   : num  14.2 13.2 13.2 14.4 13.2 ...
 $ V2   : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
 $ V3   : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
 $ V4   : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ V5   : int  127 100 101 113 118 112 96 121 97 98 ...
 $ V6   : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ V7   : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
 $ V8   : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
 $ V9   : num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
 $ V10  : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
 $ V11  : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
 $ V12  : num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
 $ V13  : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

Fig. 5: Structure of the Variables.

Table 3: List of the Variables

Original data variable	Renamed variable
V1	Alcohol
V2	Malic Acid
V3	Ash
V4	Alkali of Ash
V5	Magnesium
V6	Total of Phenol
V7	Flavonoid
V8	Bioflavonoid Phenol
V9	Pro-Anthocyanin
V10	Degree of Color
V11	Degree of light
V12	OD280/OD315
V13	Proline

3.1. Hierarchical clustering

Next, we can see the distribution of class of clusters as in table 4,

Table 4: Number of Clusters and Data Sets

Cluster #	1	2	3
Number of Data set	59	71	48

Cluster1 has 59, cluster2 has 71, and cluster3 has 48 data sets. We will then discuss the Hierarchical Clustering analysis. We have decided using 'NbClust' function and trees by using a minimum of 2 clusters, and a maximum of 6 clusters using all the coefficients with the Complete Clustering method.

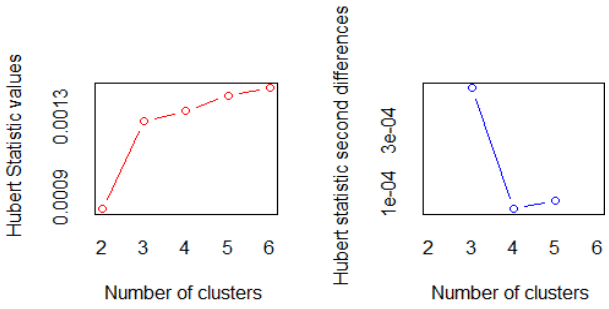

Fig. 6: Hubert Index Plot with the Complete Method.

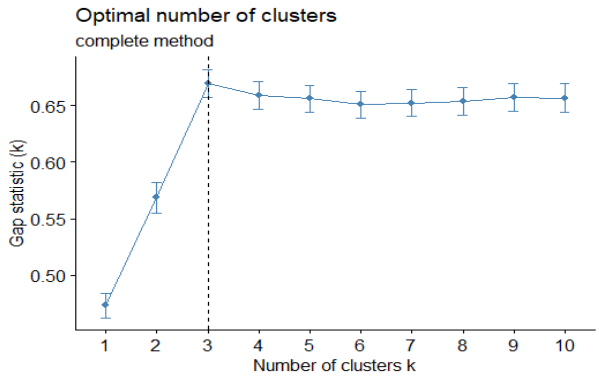

Fig. 7: Optimal Number of Clustering Using Complete.

From Figure 6 and 7, 'Hubert Index Plot', we notice that if the number of clusters is equal to 3, a big change has occurred. It means that the optimal number of clusters should be classified into three clusters using the original data set.

By using the 'Best.nc' function, we were able to calculate the number of clusters for each coefficient as in Table 5. KL, CH, Hartingan, CCC, Scott, Marriot, TrCovW TraceW, and Friedman represent the coefficients that return the optimal number of clusters to determine the distance.

Table 5: Number of Clusters and Value Index of Each Variable Using Complete Method

	KL	CH	Hartigan	CCC	Scott
Number of Clus-ters	5.00	3.00	3.00	5.00	3.00
Value_In-dex	14.2227	48.9898	27.8971	1.148	340.9634
	Marriot	TrCovW	TraceW	Fried-man	
Number of Clus-ters	3.00	3.00	3.00	3.00	
Value_In-dex	6.872	22389.83	256.4861	10.6941	

We notice that the first coefficient with 'KL' returns 5 clusters, but the 'CH' returns 3 clusters only, and so on. By these results, we created a representation graph as in the below dendrogram for the clustering.

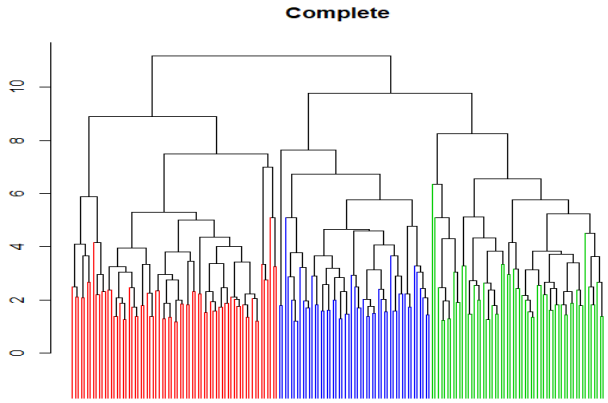

Fig. 8: Complete Cluster Dendrogram.

In Figure 8, the plots that used the complete cluster are separated into three significant clusters. Each of the data points with the same height represents the connected data sets regardless of the distance between them.

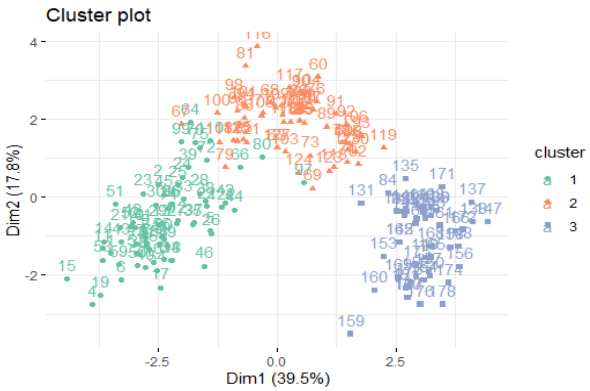

Fig. 9: Clustering with Complete Using Wine Data.

For Table 6, the data are classified into three different clusters using the Ward Cluster method.

Table 6: Number of Clusters and Data Sets Using the Complete Cluster Method

1	2	3
69	58	51

Next, we have found out the number of data sets that each cluster contains, as in Table 7.

Table 7: Number of Data Sets Each Cluster Contains Using the Complete Cluster Method

Complete	1	2	3
1	51	18	0
2	8	50	0
3	0	3	48

In table 7, each row represents the number of clusters, and each column represents the coefficients of the 'cultivar', where the cultivar represents the identification of the actual data set. We computed the difference between the Complete Cluster method results and the cultivar above. The Complete Clustering' accuracy is 84% which is calculated using the mathematical formula$\frac{51+50+48}{69+58+51} * 100\%$, where the denominator represents the total number of clusters that the data set contains. In Figures 8 and 9, we concluded that 3 is the optimal number for making different clusters. Besides that, we attach the dendrogram as below,

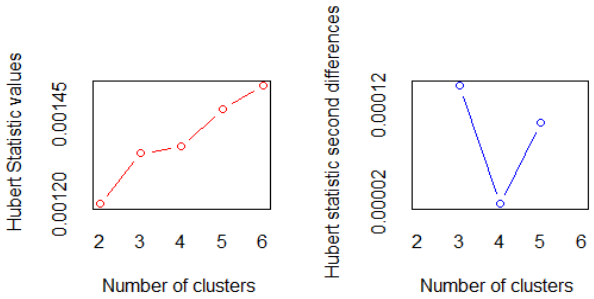

Fig. 10: Hubert Index Plot Ward Clustering Method.

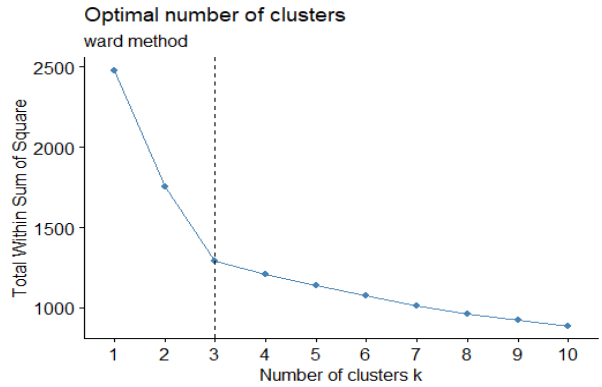

Fig. 11: Optimal Number of Clustering Using Ward Method.

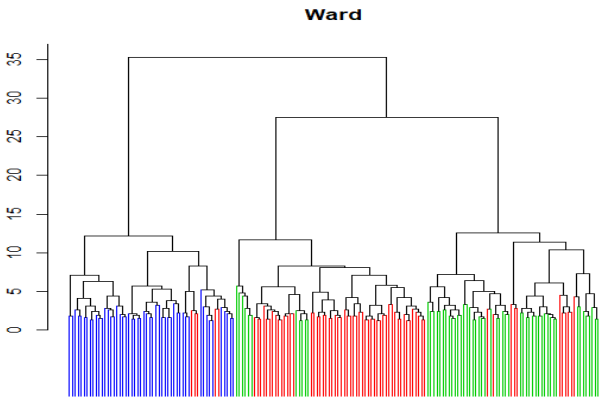

Fig. 12: Ward Cluster Dendrogram.

For Figures 10 and 11, we calculated the size of the clusters and cultivar's coefficients, where cultivar represents the identification of the actual data set. We want now to compute the difference between the Ward Cluster method results and the cultivar.

Table 8: Number of Data Sets Each Cluster Contains Using the Ward Cluster Method

Ward	1	2	3
1	59	5	0
2	0	58	0
3	0	8	48

The accuracy value of the model in the above table is 93%, which was calculated using the mathematical formula $\frac{59+58+48}{64+58+56} * 100\%$. It gave a higher accuracy value than using the complete clustering method. It also produced a higher accuracy value compared to the Ward Clustering method.

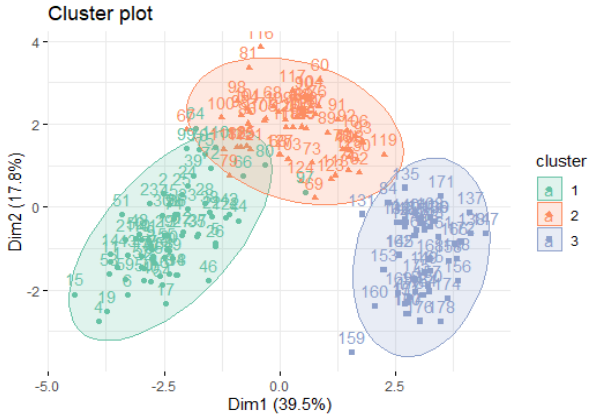

Fig. 13: Clustering with Ward Using Wine Data.

3.2. K-means clustering

By using the same 'NbClust' function to build the K-means method, we set the maximum number of clusters to 15. For Figure 14 and 15, we can see that the optimal number of clusters is 3, the same as we did with the Hierarchical Clustering method since there exists the extreme increasing and decreasing Hubert statistics value.

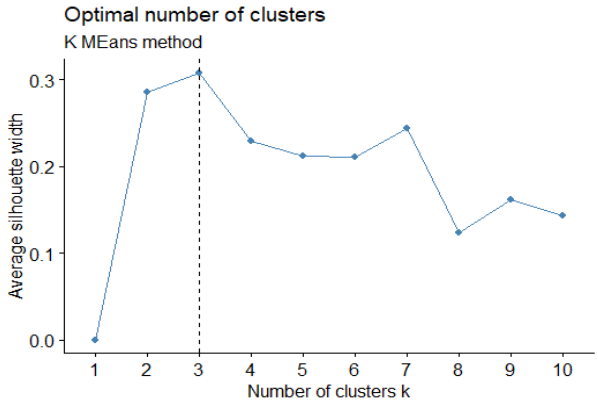

Fig. 14: Optimal Number of Clustering Using K-Means.

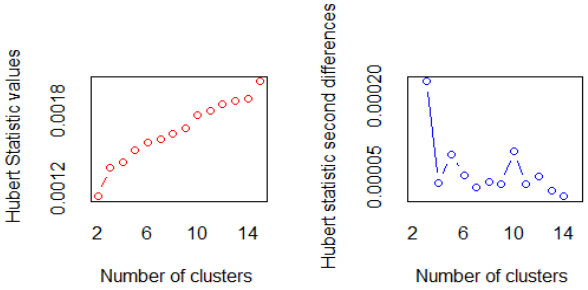

Fig. 15: Hubert Index Plot Using K Means Cluster Method.

Table 9: K-Means Value of the Total Variables. the Rows Represent the Cluster Groups, And Columns Represent the Variable of Data Set

Var-ia-ble	Alcohol	MalicAcid	Ash	Alk_ash	Magnesium	
1	0.832886	0.3029551	0.3636801	0.6084749	0.57596208	
2	0.9234669	0.3929331	0.4931257	0.1701220	0.4902869	
3	0.1644436	0.8690964	0.1863726	0.522824	0.0752647	
var-ia-bles	T_phenols	Flavanoids	Non_flav	Proantho	C_Intensity	
1	0.88274724	0.975069	0.5605853	0.57865427	0.1705823	
2	0.0757691	0.02075402	0.0343924	0.8993770	0.899377	
3	0.9765548	1.2182921	0.72402116	0.4605046	0.9388902	
var-ia-bles	Hue	0D280_315	Proline			
1	0.4726504	0.7770551	1.1220202			
2	0.4605046	0.2700025	0.7517257			
3	1.1615122	1.2887761	0.4059428			

As in Table 9, the variable 'Alcohol' has the higher K-means average value among the variables in this set. Therefore, we test this data set and compare each cluster's 'Alcohol' value using the K-means method.

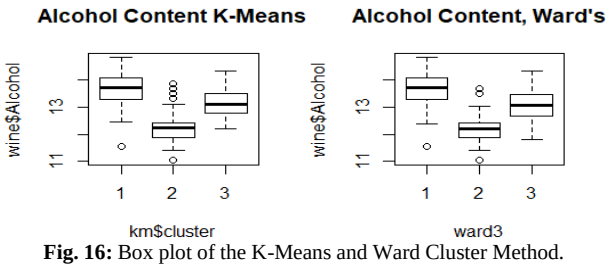

Fig. 16: Box plot of the K-Means and Ward Cluster Method.

Referring to Figure 16, when comparing the K-means method to Ward's Cluster method. Each of the clusters yields similar distribution. Thus, we conclude that using three different clusters is the best latent structure with this data set. Then, we calculate the cultivar data's coefficient value where cultivar represents the identification of the actual data set. Below, we measured the difference between the K-Means Cluster method results and the cultivar.

Table 10: Number of Clusters and Data Set Using K-Means Cluster Method

K-Means	1	2	3
1	59	3	0
2	0	65	0
3	0	3	48

The accuracy value of the model in table 10 is 97% which is calculated using the mathematical formula$\frac{59+65+48}{62+65+51} * 100\%$. As in Table 10, the result is close to the previous Hierarchical Clustering method. It concludes that by using either hierarchical or K-means clustering it produces a similar output.

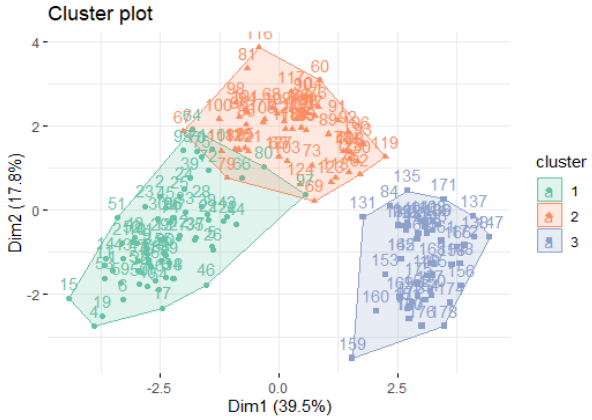

Fig. 17: Clustering with K -Means Using Wine Data.

3.3. PAM (Partition Around Medoids) clustering with Gower dissimilarity coefficient

To work on PAM method, we need to represent the data set as binary variables. We looked at the variable 'Alcohol' which is a categorical variable containing 'High' or 'Low' values. Since we changed it into the categorical variable, we have to determine the dissimilarity matrix using 'daisy' function. For figure 18, we used three different clusters previously; therefore, we also need to use 3 clusters for this method.

Table 11: Number of Clusters Using PAM Cluster Method

1	2	3
63	67	48

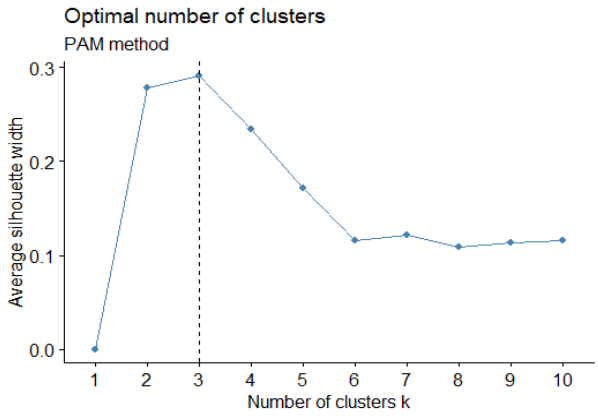

Fig. 18: Optimal Number of Clustering Using PAM.

We calculated the coefficients of cultivar, where cultivar represents the identification of the original data set. For Table 12, we measured the difference between the PAM method results and the cultivar as below. The accuracy of the model in table 12 is 94% which is calculated using the mathematical formula $\frac{57+64+47}{63+67+48} * 100\%$.

Table 12: Number of Data Sets Each Cluster Contains Using PAM Cluster Method

	1	2	3
1	57	6	0
2	2	64	1
3	0	1	47

As in Table 12, we obtained the result with descriptive statistics. In the first step, we used 'compareGroups' function to build the clusters' descriptive statistics table.

--------Summary descriptives table by 'cluster'---------

	1 N=62	2 N=71	3 N=45	p.overall
Class	1.08 (0.27)	2.04 (0.31)	2.96 (0.21)	<0.001
Alcohol:				<0.001
High	62 (100%)	1 (1.41%)	29 (64.4%)	
Low	0 (0.00%)	70 (98.6%)	16 (35.6%)	
MalicAcid	2.00 (0.83)	1.95 (0.91)	3.41 (1.09)	<0.001
Ash	2.42 (0.27)	2.28 (0.30)	2.44 (0.18)	0.002
Alk_ash	17.2 (2.75)	20.2 (3.20)	21.6 (2.26)	<0.001
magnesium	105 (11.7)	95.6 (16.8)	99.1 (10.8)	0.001
T_phenols	2.83 (0.36)	2.20 (0.56)	1.71 (0.37)	<0.001
Flavanoids	2.96 (0.42)	1.99 (0.76)	0.81 (0.31)	<0.001
Non_flav	0.29 (0.07)	0.36 (0.12)	0.46 (0.12)	<0.001
Proantho	1.89 (0.43)	1.60 (0.60)	1.18 (0.43)	<0.001
C_Intensity	5.44 (1.29)	3.17 (1.01)	7.51 (2.36)	<0.001
Hue	1.07 (0.13)	1.03 (0.21)	0.69 (0.13)	<0.001
OD280_315	3.12 (0.36)	2.75 (0.58)	1.70 (0.26)	<0.001
Proline	1070 (279)	535 (167)	635 (119)	<0.001

Fig. 19: Summary of A Descriptive Table with 3 Clusters Using PAM Cluster Method

Figure 19 lists that factor distribution, coefficient of average, and variance.

3.4. Random Forest and PAM clustering

To test the random forest with the PAM clustering method, we used 2,000 different trees for the same data set.

```
Call:
 randomForest(x = wine[, -1], ntree = 2000, proximity = TRUE)

               Type of random forest: unsupervised
                     Number of trees: 2000
No. of variables tried at each split: 3

          1         2         3         4          5
1 1.0000000 0.27598566 0.4111111 0.3593220 0.17454545
2 0.2759857 1.00000000 0.1923077 0.1757812 0.04104478
3 0.4111111 0.19230769 1.0000000 0.3694030 0.23577236
4 0.3593220 0.17578125 0.3694030 1.0000000 0.10909091
5 0.1745455 0.04104478 0.2357724 0.1090909 1.00000000
```

Fig. 20: Variable Impotence Matrix Using 2,000 Random Forest with PAM.

Figure 20 shows a small snippet of the variable importance matrix (2000 x 2000). The value of matrix elements is the probability of going from a corresponding row to a column of which data set comes together. As in Table 13, the variable 'Alcohol' can be deleted from the given data set, since it has the smallest value of 'Mean Decreasing Gini' parameter.

Table 13: Value of the Mean Decreasing Gini Parameter

	MeanDecreaseGini
Alcohol	0.5614071
MalicAcid	6.8422540
Ash	6.4693717
Alk_ash	5.9103567
Magnesium	5.9426505
T_phenols	6.2928709
Flavanoids	6.2902370
Non_flav	5.7312940
Proantho	6.2657613
C_Intensity	6.5375605
Hue	6.3297808
0D280_315	6.4894731
Proline	6.6105274

For Table 14, we calculated the dissimilarity matrix using the formula $\sqrt{(1 - \text{proximity})}$, where the value of the proximity is estimated from random forest.

Table 14: Dissimilarity Matrix Values

	1	2
1	0	0.8605821
2	0.8605821	0

We measured the difference between the Random Forest method results and the cultivar as below.

Table 15: Number of Data Sets Each Cluster Contains Using Random Forest

Random For-est	1	2	3
1	55	4	0
2	5	64	2
3	0	6	42

In Table 15, the number of clusters at each column represents the 'cultivar' coefficients where cultivar represents the identification of the actual dataset. In this case, the Random Forest accuracy is 90%, which is calculated as $\frac{55+64+42}{59+71+48} * 100\%$ where the denominator represents the total number of clusters data set contains.

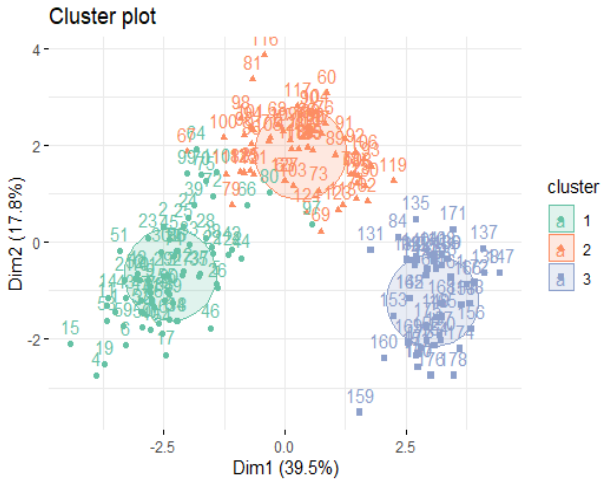

Fig. 21: Clustering with Random Forest Using Wine Data.

Table 16: Accuracy of the Model Using Hierarchical, K-Means, and PAM, and Random Forest.

	Com-plete	Ward	K-Means	PAM	Random Forest
Accuracy value	84%	93%	97%	94%	90%

4. Conclusion

In conclusion, we tested four different clustering methods, Hierarchical Clustering, K-means, PAM clustering, and Gower with Random Forest. First, we compared each of the accuracy values that obtained by using three different clusters. We used numerical and categorical variables to perform the analysis. We also utilized numerical variables and directly tested the data set using Hierarchical Clustering and the K-means method. We had to convert the data set into the categorical variable by using PAM clustering and Gower with the Random Forest method.

We obtained similar results in some cases, such as when we used Hierarchical Clustering or the K-means method when the 'Alcohol' variable is numerical. In contrast, when we changed the 'Alcohol' into the categorical variable ('High' or' Low'), we got a slightly different value when using the Hierarchical Clustering and K-means methods. When we used Ward's Clustering method, which is based on the Hierarchical approach, we obtained the highest accuracy value compared to the rest of the methods.

5. Future work

For the future work, we will focus on the Persistence Homology, Hierarchical Method, and K-means to compare the clustering given categorical data. Since we concentrate on the Persistence Homology based on the Topology and other primary clustering methods for this research, we analyzed the numerical variable data to compare each clustering algorithm to find which one is more proper among them. Next, we can test the CLARA and CLARANS clustering methods, which are based on the Machine Learning tools. Additionally, we can analyze other data, which contains the multiple factored or categorical variables to consider which clustering method is appropriate for comparing the numerical variables data.

References

[1] G. Carlsson, "Topology and data". In: Bulletin of the American Mathematical Society 46.2 (2009) pp. 255-308. https://doi.org/10.1090/S0273-0979-09-01249-X.
[2] F. Chaze, V. de Silva, M. Glisse, and S.Y. Oudot. The Structure an stability of persistence modules. Research Report arXiv:1207.3674 [math.AT] To appear as volume of SpringerBriefs in Mathematics. 2012.
[3] K. Meeham, D. Meyer. "Interleaving distance as a limit". arXiv:1710.11489v1 [math.AT] 2017.
[4] K. Meeham, D. Meyer. "An isometry theorem for generalized persistence modules". arXiv:1710.02858v1 [math.AT] 2017.
[5] S.Y. Oudot. Persistence Thoery: From Quiver Representations to Data Analysis. American Mathematical Society, 2015. https://doi.org/10.1090/surv/209.
[6] Gao, Jing. "Clustering Lecture 3: Hierarchical Methods." *Clustering Lecture 3: Hierarchical Methods*, 2019, cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_hierarchical.pdf.
[7] Topological Data Analysis gen_feedback_link (leftr, rightr); (2016, July 25). Retrieved from https://researcher.watson.ibm.com/researcher/view_group.php?id=6585.
[8] Alaa, H. N., & Mohamed, S. A. (2017, July 24). On the Topological Data Analysis extensions and comparisons. Retrieved from https://www.sciencedirect.com/science/article/pii/S1110256X17300433.
[9] Herbert Edelsbrunner and John Harer. Persistent Homology – a Survey [PDF fiwle]. Retrieved from https://www.maths.ed.ac.uk/~v1ranick/papers/edelhare.pdf.
[10] Herbert Edelsbrunner∗ and Dmitriy Morozov†. Persistent Homology: Theory and Practice. Retrieved from https://pdfs.semanticscholar.org/cf6d/43b39d66a6c3f061afeb73327312ca9cc4cb.pdf.
[11] Peter Bubenik University of Florida Department of Mathematics. Topology for Data Science 1: An Introduction to Topological Data Analysis. https://people.clas.ufl.edu/peterbubenik/files/abacus_1.pdf.
[12] Peter Bubenik, Department of Mathematics Cleveland State University. Statistical Topological Data Analysis using Persistence Landscapes. http://www.jmlr.org/papers/volume16/bubenik15a/bubenik15a.pdf.

[13] Anon, (2019). [online] Available at: https://www.quora.com/What-are-the-most-relevant-findings-and-limitations-of-Topological-Data-Analysis [Accessed 25 Oct. 2019].
[14] k-Means Advantages and Disadvantages Clustering in Machine Learning. (n.d.). Retrieved from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages.
[15] Marina Santini, Department of Linguistics and Philology Uppsals University, Advantages & Disadvantages of K-Means and Hierarchical clustering (2016), retrieved from http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf.
[16] What are the Strengths and Weaknesses of Hierarchical Clustering? (n.d.). Retrieved from https://www.displayr.com/strengths-weaknesses-hierarchical-clustering/.
[17] Hierarchical clustering algorithm - Data Clustering Algorithms. (n.d.). Retrieved from https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm.
[18] K-Means Advantages and Disadvantages | Clustering in Machine Learning. (n.d.). Retrieved from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages.
[19] Marina Santini, Department of Linguistics and Philology Uppsals University, Advantages & Disadvantages of K-Means and Hierarchical clustering (2016), retrieved from http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf.
[20] Unknown. (1970, January 1). K-Means Clustering Advantages and Disadvantages. Retrieved from http://playwidtech.blogspot.com/2013/02/k-means-clustering-advantages-and.html.
[21] Keppel, J., & Schmalz, S. (2017, November 27). Anomaly Detection: (Dis-)advantages of k-means clustering - inovex-Blog. Retrieved from https://www.inovex.de/blog/disadvantages-of-k-means-clustering/.
[22] B. Rieck1,2 and H. Leitte1, exploring and comparing clustering's of multivariate data sets using persistent homology, file:///E:/2019%20Project/reasearch3.pdf.