Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 10, 2022, page 21 - 28

Classification of Food Composition Using Agglomerative Hierarchical Clustering Method

Amirul Akram Ab. Aziz, Norhaiza Ahmad*
Department of Mathematical Sciences, Faculty of Science, University Teknologi Malaysia
*Corresponding author: norhaiza@utm.my

Abstract Food composition databases are used for a variety of objectives, including developing standards, nutrition labelling, diet and illness research, teaching, and assisting consumers in making healthier food choices. With the rising prevalence of diet-related chronic diseases such as diabetes, information on sugar content in foods is required to allow for product reformulation and to improve the effectiveness of nutritional advice. The purpose of this study was to identify food type characteristics and nutrients in the Australian Food Composition Database. Therefore, we apply agglomerative hierarchical clustering method using Complete linkage, Average linkage, Single linkage and Ward method to determine the classification of types of foods and nutrients. As a result, we have found that clustering the types of food and group of nutrients in each food consist of similar types food and several compositions of nutrients.

Keywords Food Composition, Cluster Analysis, Agglomerative Hierarchical Cluster.

1 Introduction

1.1 Food Composition

Unhealthy diets are a major risk factor for the most common non-communicable diseases, including cardiovascular disease, diabetes, and cancer. According to the Global burden of disease 2017, unhealthy diets were responsible for 42% of deaths to cardiovascular disease, 32% to type 2 diabetes mellitus, 7% to cancer, and for 16% of all deaths among adults in Western Europe (Afshin et al., 2019). The study of the chemical composition of foods is necessary for the development of nutrition science; knowledge of this field is important for professionals who work in nutrition, food data composition, food security planning, noncommunicable disease prevention, and healthcare in general. Food collection and analysis to provide data for food composition databases is very expensive, and there are competing demands for available resources to analyse new foods and newer food components, as well as to reanalyse foods with older data.

1.2 Problem Statement

The purpose of this study is to cluster various types of foods and nutrients at the same time using the agglomerative hierarchical clustering method. The procedure is designed to equalize the statistical significance of all variables used. The mathematical difficulties involved in these calculations have now been eliminated, thanks to statistical software packages of a broad scope that are also easy to use, such as R programming and Microsoft Excel employed in this work.

1.3 Research Objectives

The objectives of the research are: To identify similar characteristics of types of food and nutrients in the Australian Food Composition Database using agglomerative hierarchical clustering.

1.4 Significance of the study

Due to the increasing cases of abnormal eating disorders, it may provide quantitative information of the food composition pattern, which may assist clinicians and researchers in conducting some research. The discussion's conclusion may be able to broaden the application of the hierarchical clustering method to other sectors and highlight the advantages of the hierarchical clustering method as one of the most efficient methods for performing cluster analysis.

1.5 Scope of the study

To verify and grouping the compositions of nutrients in different types of food, Australian Food Composition Database was chosen. The database was compiled on their website which is Food Standard Australian New Zealand (FSANZ). This study will focus on details about agglomerative hierarchical clustering methods.

2 Literature Review

2.1 Introduction

Several studies and researchers related to cluster analysis and hierarchical clustering method have been discussed and attached here.

2.2 Cluster Analysis

Cluster analysis refers to a variety of algorithms and methods for categorising objects that are similar in nature.

2.3 Definition

Cluster analysis as the classification of similar objects into groups, where the number of groups as well as their forms is unknown. A similar definition is given by Everitt who studied about deriving a useful division into a number of classes, where both the number of classes and the properties of the classes are to be determined. Cluster analysis is a class of statistical techniques that can be applied to data that exhibit “natural” grouping. Cluster analysis sorts through the raw data and groups them into clusters.

2.4 Types of Cluster Analysis

Clustering itself can be categorized into 2 types of cluster analysis methods. On the basis of categorization of dataset into a particular cluster, cluster analysis can be divided into 2 types - hard and soft clustering. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters. The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions which is Centroid-based Clustering, Distribution-based Clustering and Density-based Clustering.

2.5 Objectives

The goal of this technique is to divide a set of objects, such as variables or individuals, that are characterised by a number of attributes into a set of clusters or classes, so that the objects in a class are maximally similar to each other and maximally different from each other, using a predetermined list of descriptive indicators and characteristics as the basis of the analysis.

2.6 Terminology

Cluster analysis has gone by several names, including numerical taxonomy, automatic classification, bryology, and typological analysis. Cluster analysis is known as "Numerical Taxonomy," also known as "Typology" in biology, ecology, and botany.

2.7 Mechanism and Applications

Cluster analysis identifies and classifies objects, individuals or variables on the basis of the similarity of the characteristics they posses. Dividing customers into homogeneous groups is one of the basic strategies of marketing. A market researcher may ask how to group consumers who seek similar benefits from a product so they can communicate with them better.

3 Research Methodology

3.1 Introduction

In a hierarchical classification the data are not partitioned into a particular number of classes or clusters at a single step. (Everitt et al., n.d.) Instead, the classification is made up of a series of partitions that can range from a single cluster containing all individuals to n clusters each containing one individual. It is not necessary to pre-determine the number of clusters in hierarchical clustering, as is the case with partitioning clustering.

3.2 Agglomerative Hierarchical Clustering Techniques

Hierarchical clustering techniques are divided into agglomerative methods, which involve a series of successive fusions of the n individuals into groups, and divisive methods, which involve successively separating the n individuals into finer groupings. Each observation is initially considered as a cluster of its own leaf in agglomerative clustering. The most similar clusters are then merged one by one until there is only one large cluster. Hierarchical clustering produces a tree-based representation of the objects, also known as a dendrogram.

3.3 Agglomerative Hierarchical Clustering Methods

There are many methods based on AHC, but they all differ in how the distance is defined.

- Single Linkage: This method defines the distance between two clusters based on the minimum distance of each pair of members from the two clusters. According to this definition, two clusters in each step are combined with the smallest linkage distance.
- Average Linkage: This method defined the distance between two clusters based on the average distance between all members in these clusters. According to this definition, two clusters in each step are combined with the smallest average linkage distance.
- Centroid linkage: This method defines the distance between two clusters based on the average vector distance of these clusters. According to this definition, two clusters in each step are combined with the smallest center distance.
- Complete Linkage: This method defines the distance between two clusters based on the maximum distance between all members in these clusters. According to this definition, two clusters in each step are combined with the smallest complete linkage distance.

3.4 Hierarchical Algorithms and Steps

To use agglomerative hierarchical clustering with R software, follow the steps outlined below.

1. Data preparation is performed.
2. Calculating similarity information for each pair of objects in the dataset.
3. Using the linkage function, create a hierarchical cluster tree out of the objects in step 1 based on distance. Objects/clusters that are in closed proximity are linked together using the linkage function.
4. Choosing where to divide the hierarchical tree into clusters. This creates a data partition.

3.5 Description of Distance and Similarity

Cluster analysis relies heavily on distance and similarities. Almost all clustering techniques involve determining the distance between two data points or two clusters, or the magnitude of their similarity, in order to determine which items must or must not be grouped together. Using the similarity coefficients of two data points to determine their similarity; the higher the similarity coefficient, the more similar the objects.

4 Result and Discussion

4.1 About Australian Food Composition Database

The most important aspect of the data selection process is to ensure that all of the data has a similar weight. This is done to ensure that no errors occur during the interpretation of data and results. The data collected from the website is 1616 types of food in 12 categories of nutrients only. There are some several data that contain unfamiliar and zero values.

4.2 Result

For zero bound data, standardizing the data by subtracting the mean and dividing by the zero standard deviation may be inappropriate. As a result, Table 4.3.1 displays the data standardization obtained by the R function scale ().

Table 4.2.1: Extraction of standardized data

	Food1	Food2	Food3	Food4	Food5
V1	0.55534	0.6179	0.2252	0.7730	0.7089
V2	0.2703	0.2560	-0.3510	034224	0.2660
V3	-1.6658	-1.5806	-1.5874	-1.6113	-1.6454
V4	-0.1453	0.0851	-0.7482	-0.5709	-0.0034
V5	-0.1606	0.0698	-0.7535	-0.5778	-0.0179
V6	0.7663	1.9377	0.3367	0.7272	0.8053
V7	2.2130	1.8522	6.4267	3.9080	4.4322
V8	3.6224	2.1192	0.4469	4.1861	5.5201
V9	0.11564	0.9732	-0.6857	-0.4045	1.7394
V10	0.5695	1.1851	0.0598	0.4961	1.6791
V11	-0.1655	1.5129	-0.1702	-0.0129	-0.1555
V12	2.6688	1.1923	0.0419	0.2701	1.3786

The mean and standard deviation of the types of food and nutrients are shown in Table 4.3.2. As can be seen in Table 4.3.2, descriptive statistics on food composition are summarised. Energy with dietary fibre has the highest mean value for all variables which is 845.64.

Table 4.2.2: Means and Standard Variation of Food Composition

	Mean	Standard Deviation
"Energy with dietary fibre, equated (kJ)"	845.64	702.91
"Energy, without dietary fibre, equated (kJ)"	823.60	696.85
"Moisture (water) (g)"	57.14	29.32
"Protein (g)"	12.44	11.28
"Nitrogen (g)"	2.01	1.82
"Ash (g)"	1.87	5.12
"Calcium (Ca) (mg)"	57.91	146.9
"Magnesium (Mg) (mg)"	36.21	53.22
"Phosphorus (P) (mg)"	161.55	142.26

"Potassium (K) (mg)"	350.15	1349.96
"Sodium (Na) (mg)"	298.99	1697.33
"Zinc (Zn) (mg)"	1.74	2.15

From the figure below, the Potassium (K) and Sodium (Na) have some outlier from the data given compared to other. It's important to investigate the nature of the outlier before deciding. Outliers could be the result of chance or they could indicate something scientifically interesting.

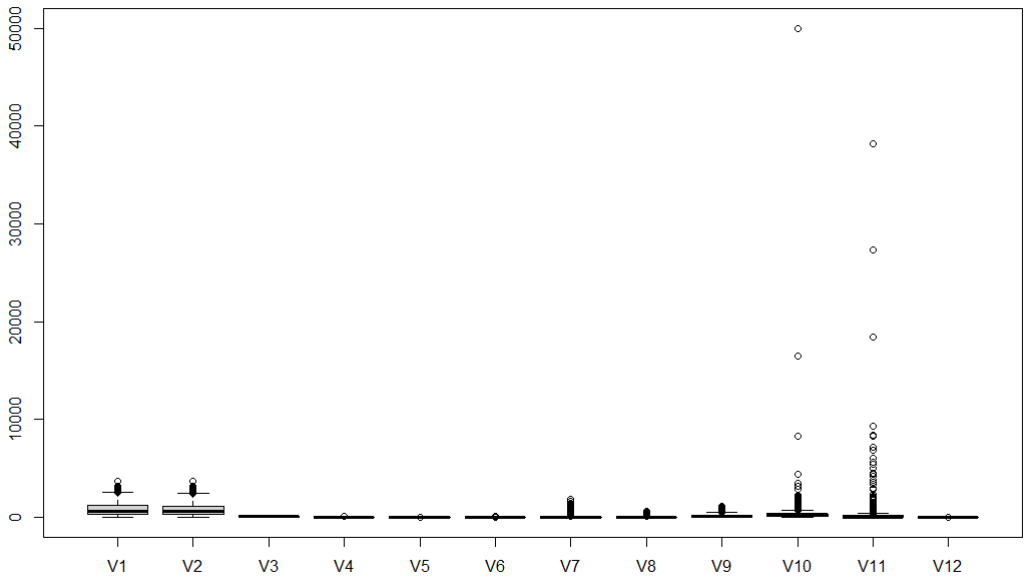

Figure 4.2.1: The boxplot of food composition

4.3 Distance Measure and Agglomerative Hierarchical Clustering Results

Distance is one of the most important roles in cluster analysis which is used to describe the quantitatively about dissimilarity of two data points or how dissimilar two clusters are. Two data points or two clusters are said to be more dissimilar when the greater dissimilarity measure is obtained. Cluster analysis will group together objects that have many similarities and separate objects that differ from one another.

Table 4.3.1: Table contained similarity matrix for nutrient

	V1	V2	V3	V4	V5
V1	0.0000	7.3805	86.9454	96.2393	97.7656
V2	7.3805	0.0000	84.6711	93.1786	94.7241
V3	86.9454	84.6711	0.0000	27.6751	28.3192
V4	96.2393	93.1786	27.6751	0.0000	2.0818
V5	97.7656	94.7241	28.3192	2.0818	0.0000

Table 4.3.2: Table contained similarity matrix for types of food

	Food1	Food2	Food3	Food4	Food5
Food1	0.0000	3.1608	6.0862	3.1102	3.5924
Food2	3.1608	0.0000	6.0296	4.0519	4.8150
Food3	6.0862	6.0296	0.0000	4.6674	6.3008
Food4	3.1102	4.0519	4.6674	0.0000	2.9334
Food5	3.5924	4.8160	6.3008	2.9334	0.0000

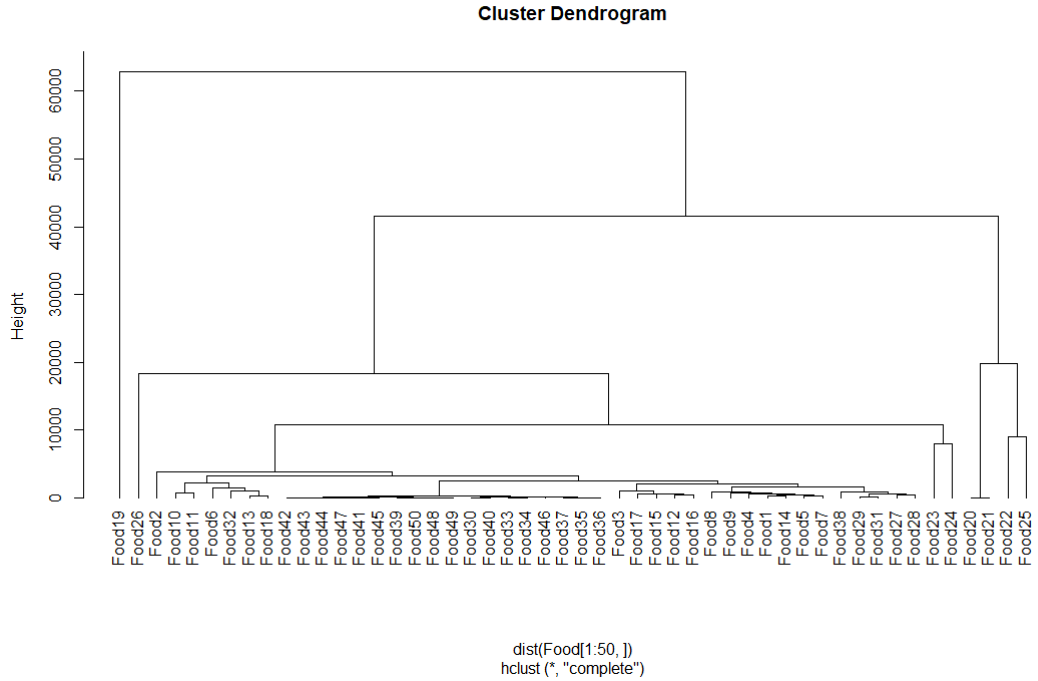

Figure 4.3.2.1: Dendrogram of 50 types of food using Euclidean distance and Complete-Linkage clustering method.

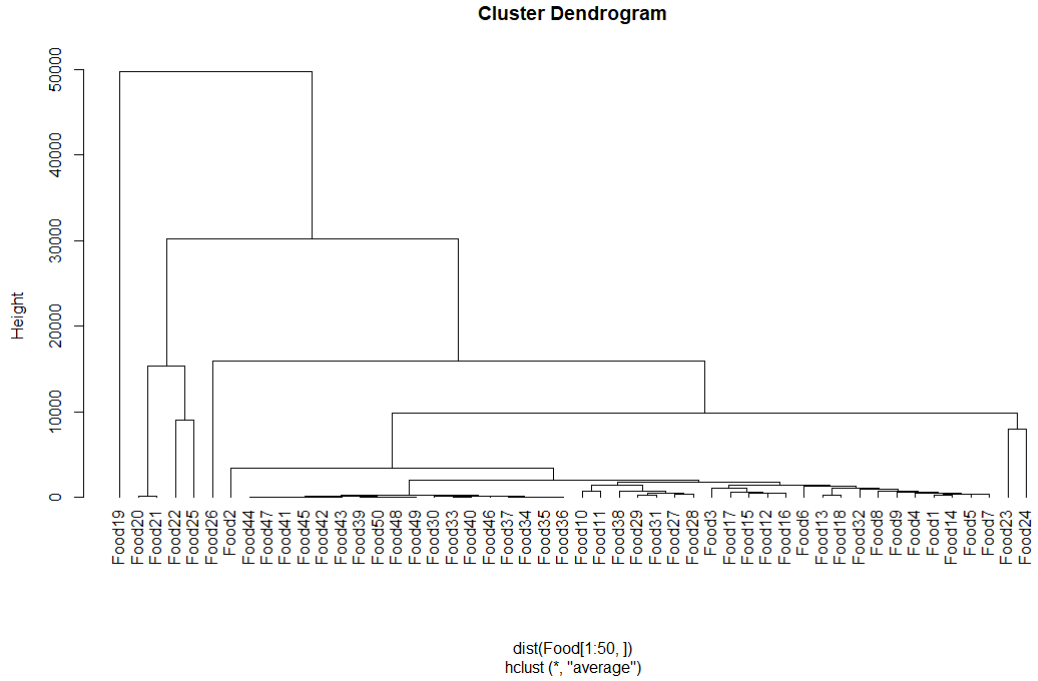

Figure 4.3.2.2: Dendrogram of 50 types of food using Euclidean distance and Average-Linkage clustering method.

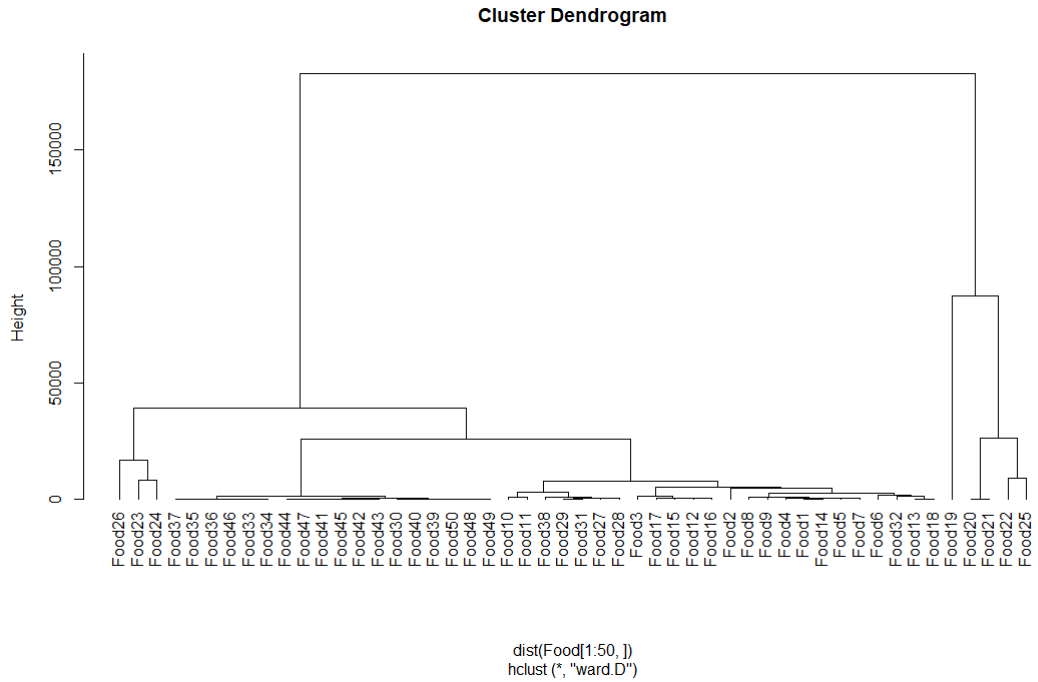

Figure 4.3.2.3: Dendrogram of 50 types of food using Euclidean distance and Ward-Linkage clustering method.

4.4 Discussion

After applying Euclidean distance and clustering method, it is conclude that the types of food, Food19- is the most dissimilar station from other food. There are 5 clusters exist in the grouping of food composition in the dendrogram. There numbers of type of food clustering can be summarised in Table as shown.

Table 4.4.1: Number nutrients in each cluster

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
V1, V2	V3, V9	V4, V5, V6, V7, V8, V12	V10	V11

Table 4.4.2: Number of types of food in each cluster

Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
184	5	295	563	569

5 Conclusion and Recommendations

5.1 Introduction

This chapter discusses about the conclusion, limitations and suggestions for future study.

5.2 Conclusion

To summaries, the objective of this study is to cluster types of food and nutrients simultaneously using agglomerative hierarchical clustering method identified. The results obtained that Food22 Food24 and Food25 groups together because it is categorized as dry powder.

Two important issues that arise along this study is the issue on data with high variability and data with high in dimension. Therefore, to minimize the variability, the data should do normalization process. The preprocessing data that carried successfully reduce some of the data that incomplete.

By using agglomerative hierarchical clustering method, the types of food and nutrients can be clustering smoothly.

5.3 Recommendations

The recommendations made during this study in dealing with the outlier. The common method of normalisation, normalisation by means, cannot suppress the problem's outliers. Thus, the first method is to eliminate outliers immediately after normalisation, but this method loses some information. The second method is to go through another normalisation process, such as median normalisation. To be more realistic, additional studies should be conducted with a larger number of variables.

6 References

[1] Burlingame, B. a. (2009). Food composition is fundamental to the cross-cutting initiative on biodiversity for food and nutrition. *Journal of food composition and analysis*, 361-365.
[2] Chapman, J. a. (2020). Application of cluster analysis in food science and technology. *Mathematical and Statistical Applications in Food Engineering*, 68-73.
[3] da Silva Torres, E. A. (2006). The application of hierarchical clusters analysis to the study of the composition of foods. *Food chemistry*, 622-629.
[4] Elmadfa, I. a. (2010). Importance of food composition data to nutrition and public health. *European journal of clinical nutrition*, S4-S7.
[5] Eminagaoglu, M. a. (2022). Evaluation of elemental affinities in coal using agglomerative hierarchical clustering algorithm: A case study in a thick and mineable coal seam (kM2) from Soma Basin (W. Turkey). *International Journal of Coal Geology*, 104045.
[6] Food composition at present: new challenges. (2019). *Kapsokefalou, Maria and Roe, Mark and Turrini, Aida and Costa, Helena S and Martinez-Victoria, Emilio and Marletta, Luisa and Berry, Rachel and Finglas, Paul*, 1714.
[7] Granato, D. a. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science \& Technology*, 83-90.
[8] Hulshof, P. a. (2019). Food composition tables in Southeast Asia: the contribution of the SMILING Project. *Maternal and child health journal*, 46-54.
[9] Li, J. a. (2019). Innovation clusters revisited: On dimensions of agglomeration, institution, and built-environment. *Sustainability*, 3338.
[10] Li, T. a. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University-Computer and Information Sciences*.
[11] Liu, N. a.-J. (2021). An agglomerative hierarchical clustering algorithm for linear ordinal rankings. *Information Sciences*, 170-193.
[12] Merchant, A. T. (2006). Food composition database development for between country comparisons. *Nutrition journal*, 1-8.
[13] Murtagh, F. a. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 86-97.