Clustering and Interpretation on Real Nutritional Data

Beatriz Sevilla Villanueva[1], Karina Gibert[1,2], and Miquel Sànchez-Marrè[1]
bsevilla@lsi.upc.edu, karina.gibert@upc.edu, miquel@lsi.upc.edu

[1] Knowledge Engineering and Machine Learning Group (KEMLg), Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, Spain
[2] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya - BarcelonaTech, Barcelona, Spain

Abstract. Nutritional Genomics studies diet-gene-disease interactions and aims to promote health and disease prevention. It is based on the idea that everything ingested into a person's body affects the genome of the individual and, therefore, both genes and nutrients modify the same metabolic processes. This paper presents an application of clustering and interpretation over real heterogeneous data coming from a nutritional study. The individuals are clustered by their diet and physical activity habits and the resulting clustering is interpreted. This work is part of a methodology to deal with data from dietary intervention studies.

1 Introduction

Nutrition has an active role in the people's health. In last decade the efforts of nutrition are oriented to the personalized diet. They wonder *Why apparently similar persons have different responses to the same nutrient?* For answering this question several clinical studies have been performed. In addition with the enormous advances in genomics, those studies can include genetic information to obtain a new view of why nutrients affect persons in a different way [31].

This field is emerging as important new research area. The reason is that it is becoming increasingly evident that damage to genome is the most fundamental disease. A risk for developing a disease increases with DNA damage, which is dependent on nutritional status and an optimal dietary intake. These genes are involved directly or indirectly in the uptake and metabolism of nutrients [9].

In this work we will analyze part of the data coming from a nutritional genomic trial. This trial is a randomized dietary intervention study for relating the gene expression with the Mediterranean diet. This study collects different information such as biometric, sociodemographic characteristics, as well as clinical test. Also, it includes a survey about their habits both diet and physical activity. The aim of this work is to cluster the individuals depending on these habits. This is a step of a methodology which is being defined for finding the diet effect, but locally to each one of the profiles of persons, in such a way that the complex interactions can be decomposed.

The data set of diet and physical activity habits is heterogeneous. Therefore, the clustering has to be adapted to this type of data. In this work, we use a hierarchical clustering using the Gower's metric to handle both numerical and categorical attributes.

A good characterization and interpretation of the resulting clusters is very important in this kind of applications. In this paper, the interpretation of these clusters is done by means of a graphical representation of the attributes against the clusters.

2 Related Work

Advances in Nutritional Genomics are being enormous in last years. Specially since the gene expression can be translated into microarrays formats which can be afterwards analyzed. Thus, this genomic data can be involved in studies such as the dietary intervention pre-post studies. Notwithstanding, there is still a long way before a highly personalized diet can be reached [31].

The publications in nutritional genomics are multiple but not all reliable as it is explained in [27]. In this review of the gene-environmental factor interactions, it is reported that in many studies there is a lack of statistical significance, many published results are contradictory, and in addition, many studies are not reproducible.

In [14] is pointed out a need for new methodologies to analyze the efficacy of interventions when contextual factors that can interact with the intervention itself are not all well-known. These conditions make difficult to design a correct pre-post studies as they were initially conceived and decrease the reliability of the provided results. The traditional techniques used in these studies not always extract all relevant information from data.

In the particular field of nutrition, several nutritional intervention trials can be found. As an example, the one conducted in Linxian, China: Dysplasia and General Population Trial [43]. They tested the effect of multiple vitamins and minerals in the prevention of esophageal cancer. Basic statistical pairwise test are mainly used to conclude in these kind of studies.

Few works are found in this field using clustering. Next works are about environmental factors including diet but do not include an intervention. In [24], diet and other lifestyle factors are analyzed using k-means. In [37], individuals are clustered depending on their diet, physical activity and anthropometric characteristics. In other medical fields, some references are found about epidemiological pre-post studies which introduce some AI techniques for the analysis. In [5], the hypertension is studied in Korea with Insurance data by comparing logistic regression, decision trees and association rules. In [38], statistical models and association rules were used to study the comorbidity of attention deficit-hyperactivity disorder in Taiwan.

Within the specific field of nutritional genomics, clustering using data generated by gene expression profiling can be used to identify sub-populations of subjects that respond differently to a given diet intervention. Nevertheless, few

works are found using clustering. Working with animals, in [28] clustering was used to typify the different kind of diets of a set of cows with more o less fats.

An interesting antecedent is [40] focused on clustering a human subcutaneous adipose tissue gene expression data obtained during low-calorie diet intervention to aid in the prediction of 6-month weight loss maintenance. This work timidly points to a need of *local studies over the different groups of intervention.*

2.1 Cluster Interpretation

In many data mining processes there is still a gap between raw data mining results and effective decision-making. Postprocessing data mining results to approach them to the decision makers is crucial for an impact of the analysis on reality. Some works points to this issue [6, 17] and for the particular case of clustering interpretation and conceptualization of clusters is a key issue to this purpose. Some ideas have been developed in previous works [2, 15, 11, 12, 36]. The main assumption is that the resulting clusters are different, at least, in base to the data used to create them. Thus, one could expect that not every attribute will be of equal importance when describing the different groups (clusters). Therefore, a comparative analysis between the attributes that have similar values together in the different clusters can be conducted, which provides new insights into the complexity of the cluster description. In [4], they rank the attributes involved into the clustering with *logWorth*. This index is based on p-values which are assessed with the χ^2-Independence Test.

Efforts in cluster interpretation also aims at visualizing the attributes through the resulting clusters. In [21] the means of the standardized attributes are displayed for every cluster. Besides, *Class Panels Graphs* are introduced in [15] as a graphical representation in form of panel, containing attributes in the columns and clusters in the rows. Conditional distributions of the attributes against the clusters are shown per column. This visualization is interesting to understand better the meaning of the clusters [13] and it is used in this work.

3 Methodology proposal

Nowadays, results obtained from classical pre-post studies are based on traditional and often basic, statistical techniques. Till now, these techniques have not been expressive enough to allow the extraction of complex relationships. The level and degree of interactions between different subset of attributes is too complex, in this context, to be captured by simple data analysis or pre-post statistical testing.

This work focuses on finding the profiles of the individuals before they started the dietary intervention. This purpose belongs to the proposal of a general methodology [34] which will help to cluster the individuals properly for subsequently analyzing the pre-post effects. This classification allows to decompose the posterior intervention analysis in such a way that the effect can be studied locally to these profiles.

Most of the clinical studies refer to numerical data. However, categorical attributes may be also of importance for the analysis. Upon [1], three main strategies may be followed in front of mixed data: *i) Partitioning:* analyzing the dominant type; *ii) Converting:* all attributes to a unique type, conserving as much original information as possible; *iii) Compatibility measures:* allowing clustering on heterogeneous data matrices without transforming the attributes themselves. Main advantages of the later approach are no loss of information, no need to take previous arbitrary decisions which can bias results, allowing studies with all types of attributes together, and the analysis of interactions between attributes of different types. In the literature, several proposals are found [16]. In this work Gower [19] is used for those algorithms where it is available and Heterogeneous Euclidean-Overlap Metric (HEOM)[42] for the other ones.

For this first step, hierarchical clustering based on Ward's method is used [41]. The Ward's method aims at finding compact and distinguishable clusters. The Ward's method is widely used because it is a criterion related to the quantity of information of the clusters, and the resulting clusters can be often interpreted easily. Hierarchical clustering does not need number of clusters as input.

Criteria to cut the dendrogram is based on optimal Calisnki-Harabasz Index [3].Calinsky-Harabasz is a compromise between both the between-cluster distance and the within-cluster distance.

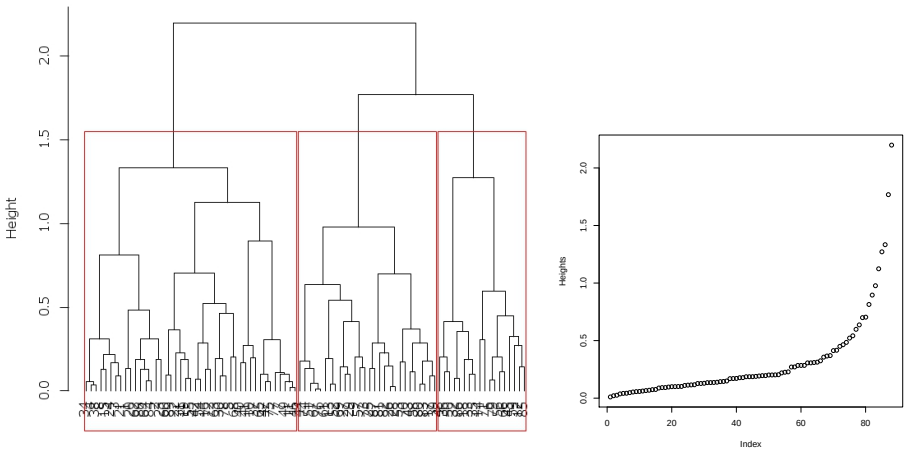
Then, the resulting partition is compared with partitions obtained by using different clustering methods, on the basis of the performance computed by 10 cluster validity indexes available in R [32]: Dunn [20], Pearson version of Hubert's gamma coefficient (Pearson)[20], average of silhouette width [33], Calinsky-Harabasz Index (CH)[3], average distance between clusters, minimum cluster separation, Separation index [23], average distance within clusters, Goodman and Kruskal's G3 index [18] and maximum cluster diameter [23].

The resulting clusters are interpreted using *Class Panel Graphs* [15] as the interpretation support tool.

4 Application & Results

As said before, the data used in this work comes from a real nutritional genomic study. Dietary habits are described through 14 categorical attributes; these attributes follow same questionnaire published in [26]. Physical activity is designed through 10 numerical attributes reporting weekly and daily activity as well as totals. Only the 4 attributes referred to the physical activity per week are used to avoid redundancies.

Therefore, the clustering is applied over these 18 selected attributes. As it was mentioned, a hierarchical clustering based on Ward's method with Gower's metrics is performed. The dendrogram of the resulting clustering is depicted in Fig. 1a. Also, in Fig. 1b, the histogram of level indexes of the dendrogram is shown. From these figures, it is possible to determine a convenient cut in 3 clusters according to max leaps in the graphics: $H = \{H_1, H_2, H_3\}$ (consistent with Calinski-Harabasz recommendation).



(a) Dendrogram cut in 3 clusters (b) Heights of level indexes

Fig. 1: Hierarchical clustering with Ward's Method for Diet and Physical Activity

Partition	Num.Clusters	Dunn	Pearson	Avg. silhouette	CH	Avg. Between	Min Separation	Separation	Avg Within	G3	Max Diameter	Partition Specifics
H	3	0.15	**0.30**	**0.13**	**13.24**	0.33	**0.09**	**0.11**	**0.27**	**0.30**	**0.58**	Ward's Method [41], Gower dist
P_1	3	0.10	**0.30**	**0.14**	**14.72**	0.33	0.06	0.07	**0.27**	**0.29**	**0.63**	PAM [25]. Gower dist, 3 clusters, selection by Calinski-Harabasz
P_2	3	**0.16**	0.27	0.12	7.04	**0.36**	**0.09**	**0.11**	0.29	0.33	**0.58**	Complete Linkage method[35], Gower dist
P_3	3	0.11	0.29	0.08	7.74	**0.34**	0.07	0.09	0.28	0.32	0.67	Model-Based Clustering [10]. 3 mixtures, model "EEE", ellipsoidal distribution, equal shapes and volumes, HOEM dist
P_4	6	**0.20**	0.29	-0.08	3.23	**0.34**	**0.13**	**0.13**	0.28	0.31	0.65	DBSCAN[8]. Epsilon:0.13, min. points:2, HOEM dist
P_5	3	0.04	0.10	0.03	4.25	0.32	0.03	0.05	0.30	0.44	0.71	k-means [29]. Max iterations: 1000, 3 clusters, HOEM dist.

Table 1: Cluster Validation Indexes. The two best values are in bold.

Table 1 describes the obtained partition, together with the results obtained with other 5 methods for comparison. R software [32] has been used for all cases, with packages: *stats, StatMatch [7], cluster[30], mclust[10], fpc[22], MASS[39].* For every partition, the 10 cluster validity indexes introduced in section 3 have been evaluated (Table 1). The first 7 indexes indicates better quality with higher values; the last 3 with lower values.

It is possible to observe from this Table 1 that H and P_1 partitions generally perform better than the rest while P_4 and P_5 have the worst index values. For instance, partition P_4 has one cluster which contains most of the instances, that is why the diameter is so high. Also, its Dunn index (based on separation) is higher because the rest of the clusters are small and isolated. Between H and P_1 small differences are observed but H improves P_1 in 4 of the 10 indexes while P_1 improves H in 3. Thus H is going to be interpreted in next section.

4.1 Cluster Interpretation

Class panel graphs are used to visualize empirical conditional distributions of attributes against clusters as it is shown in Table 2 and Table 3. Table 2 shows the 14 diet attributes and in Table 3, the original 10 physical attributes. Based on these *Class Panel Graphs* the following profiles can be described.

H_1 is a cluster that eats more vegetables but less fresh fruit. Comparing with other clusters, they consume a lower quantity of red meat, less legumes and commercial bakery. In addition, this cluster eats more fish than H_2. Their exercise is low in general, but some does from light to hard physical activity.

H_2 contains some persons that consume less oil per day. Comparing with other clusters, their consumption of red meat is higher and they use slightly more butter and takes more gas drinks. They eat less fish and nuts than other clusters. This cluster is the one which practices less physical activity.

H_3 is a cluster that eats more fresh fruit than other clusters. They eat few portions of red meat and uses less butter. This cluster eats some fish and consumes more commercial bakery and nuts than the rest of clusters. This cluster practices slightly more physical activity with a significant proportion of persons doing moderate physical activity.

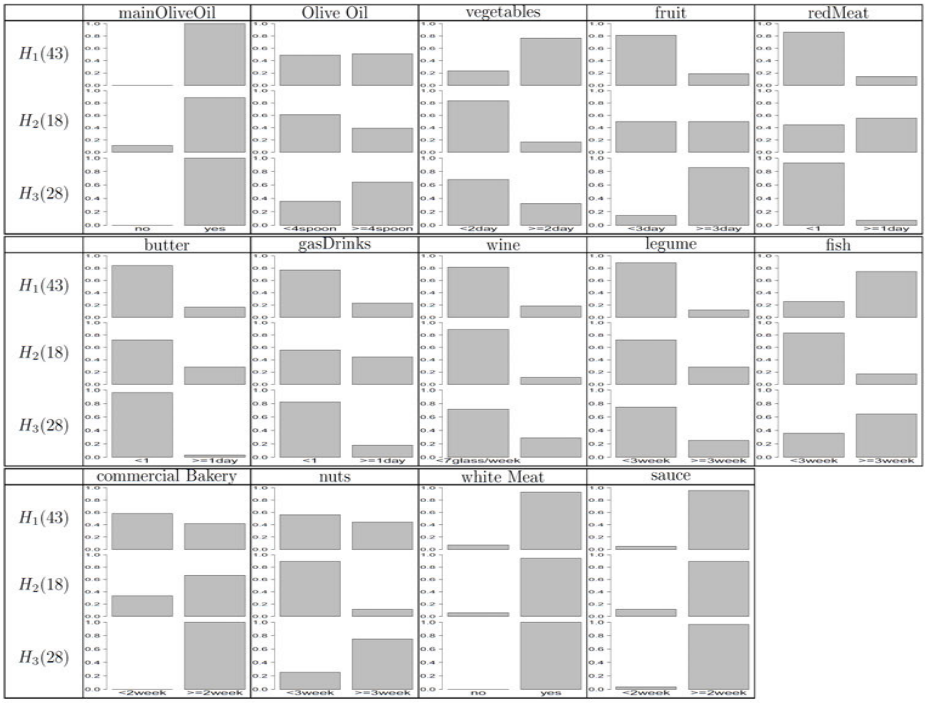


Table 2: Diet characteristics

Table 3: Physical Exercise characteristics

5 Conclusions

In this work we have introduced the drawbacks of the nutritional genomics and how AI techniques can help in extracting knowledge from data. As a first step to solve these difficulties, we have presented the classification of the individuals before the dietary intervention. Although few works are found, we are convinced that using clustering for finding patterns of response to diet may be of help to assist decision making in diet design. Therefore, clustering can bring light to this aspect because it can identify multivariate patterns in a more natural way without requiring a clear previous knowledge about the laws governing the target phenomena. Finding a prior substructure of the problem can make easier to get richer models.

Besides, the evaluation of the resulting clusters based on cluster validity indexes has been successful. Notwithstanding, the main goal achieved is the use of clustering with heterogeneous real data and the visualization to assists the interpretation of the resulting classes. Regarding cluster interpretation, not many research works are found in the literature. Methods or tools for obtaining and helping cluster interpretation seem to be a problem not yet solved.

Experts have found meaningful the composed profiles from a medical point of view. Thus, their validation of the interpretation has been positive.

Next steps will focus on finding the diet intervention effect locally to these obtaining profiles.

6 Acknowledgements

The authors acknowledge the project AGAUR (2009SGR 1365) from the catalan government oriented to support excellent research groups.

References

1. M.R. Anderberg. *Cluster Analysis for aplications.* Academic Press, 1973.
2. K. Barnard, P. Duygulu, et al. Clustering art. In *CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–434–II–441, 2001.
3. T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3:1–27, 1974.
4. W. Cecere and D.A Abreu. A method for improving list building: Cluster profiling. In *Proc. of the Survey Research Methods Section, AMSTAT*, 2010.
5. Y.M Chae, S.H Ho, et al. Data mining approach to policy analysis in a health insurance domain. *IJMI*, 62(2.3):103–111, 2001.
6. P. Cortez and M.J Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 2012.
7. M. D'Orazio. *StatMatch: Statistical Matching*, 2012. R package version 1.0.5.
8. M. Ester, H.P. Kriegel, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, KDD*, 1996.

9. M. Fenech. The human genome, nutrigenomics and nutrigenetics. *Innovation: Management, Policy & Practice*, 2008.
10. C. Fraley and A.E Raftery. Model-based clustering, discriminant analysis, and density estimation. *ASA*, 97(458):611–631, 2002.
11. K. Gibert. Automatic generation of classes interpretation as a bridge between clustering and decision making. *IJMCDM*, 2013.
12. K. Gibert, D. Conti, et al. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *EEMJ*, 11(5):931–944, 2012.
13. K. Gibert, C. García-Alonso, et al. Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. *Health research policy and systems*, 8(1):28, 2010.
14. K. Gibert and A. García. Posibilidades de aplicación de minería de datos para el descubrimiento de conocimiento a partir de la práctica clínica. *Tecnologías Aplicadas al Proceso Neurorrehabilitador*, 2008.
15. K. Gibert, A. García-Rudolph, et al. The role of kdd support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. *Acta Inform Med*, 16(4):178–182, 2008.
16. K. Gibert, R. Nonell, et al. Knowledge discovery with clustering: impact of metrics and reporting phase by using klass. *Neural Network World*, 4:319–326, 2005.
17. K Gibert, G Rodríguez-Silva, et al. Post-processing: Bridging the gap between modelling and effective decision-support. the profile assessment grid in human behaviour. *Mathematical and Computer Modelling*, 2011.
18. A.D Gordon. *Classification, 2nd Edition (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 2 edition, June 1999.
19. J.C. Gower. A General coefficient if similarity and some of its properties. *Biometrics*, 27:857–874, 1971.
20. M. Halkidi, Y Batistakis, et al. On clustering validation techniques. *JIIS*, 17:107–145, 2001.
21. D. Haughton, P. Legrand, and S. Woolford. Review of three latent class cluster analysis packages: Latent gold, polca, and mclust. *AMSTAT*, 63(1):81–91, 2009.
22. C. Hennig. *fpc: Flexible procedures for clustering*, 2013. R package version 2.1-5.
23. C. Hennig and T.F Liao. Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Technical report, Technical report, 2010.
24. K.F Hulshof, M. Wedel, et al. Clustering of dietary variables and other lifestyle factors (dutch nutritional surveillance system). *JECH*, 46(4):417–424, 1992.
25. L. Kaufman and P.J. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1–Norm and Related Methods*, pages 405–416, 1987.
26. V. Konstantinidou, M.A Covas, et al. In vivo nutrigenomic effects of virgin olive oil polyphenols within the frame of the mediterranean *FASEB J*, 24(7):2546–57, 2010.
27. Y. Lee, C. Lai, J.M. Ordovas, and L. Parnell. A database of gene-environment interactions pertaining to blood lipid traits, cardiovascular disease and type 2 diabetes. *JDMGP*, 2(1):1–8, 2011.
28. J.J Loor, M. Bionaz, et al. Systems biology and animal nutrition: insights from the dairy cow during growth and the lactation cycle. *Systems Biology and Livestock Science*, 2011.
29. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. California, USA, 1967.

30. M. Maechler, P. Rousseeuw, et al. *cluster: Cluster Analysis Basics and Extensions*, 2012. R package version 1.14.2.
31. J.M Ordovas and V. Mooser. Nutrigenetics and nutrigenomics. *Casopis Lekaru Ceskych*, 146(2):837–839, 2007.
32. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.
33. P.J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
34. B. Sevilla, M. Sànchez-Marrè, and others. Using ai in nutritional genomics. In *Poster in Sematica 2012, Granada*, November 2012.
35. R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.
36. M. Siponen, J. Vesanto, et al. An approach to automated interpretation of som. In *Advances in Self-Organising Maps*, pages 89–94. Springer, 2001.
37. S. Swaminathan, T. Thomas, et al. Clustering of diet, physical activity and overweight in parents and offspring in south india. *EJCN*, 2012.
38. Y.M Tai and H.W Chiu. Comorbidity study of adhd: Applying association rule mining (arm) to national health insurance database of taiwan. *IJMI*, 78(12):e75 – e83, 2009.
39. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
40. N. Viguerie, C. Poitou, et al. Transcriptomics applied to obesity and caloric restriction. *Biochimie*, 87(1):117 – 123, 2005.
41. J.H Ward Jr. Hierarchical grouping to optimize an objective function. *ASA*, 58(301):236–244, 1963.
42. D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.
43. X.N Zou, P.R Taylor, et al. Seasonal variation of food consumption and selected nutrient intake in linxian, a high risk area for esophageal cancer in china. *Int J Vitam Nutr Res*, 72(6):375–382, Dec 2002.