Post-Processing the *Class Panel Graphs*: Towards Understandable Patterns from Data

Beatriz SEVILLA VILLANUEVA [a,1], Karina GIBERT [a] and Miquel SÀNCHEZ-MARRÈ [a]

[a] *Knowledge Engineering and Machine Learning Group (KEMLg), UPC, Spain*

Abstract. A profiling methodology is introduced for automatic interpretation of clusters in this paper. This methodology contributes to the characterization of the resulting classes from a clustering process. Our research aims to find a concordance between the proposed methodology and the experts' description of these classes. In this work the resulting classes from a clustering of a general population sample based on their diet and physical activity habits are interpreted and compared with the experts' description of these classes by using the *Class Panel Graphs*. As a novelty, we import techniques from the multivariate analysis into the cluster interpretation process.

Keywords. Clustering, Post-processing, Interpretation, Cluster Profiling

Introduction

Nowadays, it is known that there is a relationship between nutrients and health but it is still unknown how this relationship is. In order to study this relationship, several nutritional studies have been performed.

Data included in such studies come from surveys and medical tests. In most of cases an important number of characteristics are measured. The classical approach is to consider only numerical attributes but this approach is not powerful enough to clearly find the factors associated to diet.

In this work a methodology including a strong component in cluster interpretation is proposed and applied to pattern discovery on diet and physical activity habits of a general population sample. Profiling methodology can contribute to identify standard nutritional patterns in general population. State of the Art shows an open problem to get clear interpretation of the discovered patterns that can be really useful in the later decision making processes (guidelines establishment, dietary recommendations, etc).

How to show the data mining results is an important part in real applications and specially, when results must be transferred by experts from different disciplines. Thus, a good characterization and interpretation of the clusters is very important.

[1] Corresponding Author: bea.sevilla@gmail.com

In this work, clustering techniques are used to find profiles on diet and physical activity habits, and *Class Panel Graphs* are used to support the interpreting process followed by the experts. They are proposed in [1].

Going one step further the performance of several statistical tests is analyzed to see the degree of concordance with the experts' observations in order to find objective procedures to reproduce human interpretations automatically.

The structure of the paper is the following: related work is explained in the next section. Then, the proposed methodology is explained in section 2. This procedure is applied over 3 classes resulted from a previous clustering and compared with experts' description in section 3. Finally, the conclusions of this work are collected in section 4.

1. Related Work

Cluster interpretation has an important role for presenting the results to the experts. The interpretation of clusters is also found in literature as *Cluster Profiling* [2].

Assuming that the process identifies different clusters, one could expect that at least one of the attributes used for the clustering will behave differently in one cluster giving the distinctive characteristics of the cluster. Therefore, a comparative analysis between the attributes used in the clustering can be conducted, which provides new insights into the complexity of the cluster description.

The cluster interpretation is commonly made by examining the cluster centroids, which are the attribute' average values of all objects in a certain cluster. In [2] it is stated that clusters are distinguishable only if certain attributes exhibit significantly different means in some clusters, at least from a data perspective. This significance can be assessed by comparing the clusters with independent t-tests samples or ANOVA.

In [3] the attributes involved into the clustering are ranked with *logWorth*. This index is based on p-values which are usually assessed with the χ^2-Independence Test.

Efforts in cluster interpretation also go in visualizing the conditional distribution of attributes through the resulting clusters. In [4] displays the cluster means for the standardized attributes for every group. Besides, *Class Panels Graphs* are introduced in [5] as a graphical representation in the form of panel, containing attributes in the columns and classes in the rows. Conditional distributions of the attributes against the classes are shown per column. This visualization is interesting to identify specific behaviors of some attributes in a certain class and thus to understand better the meaning of the classes [6]. *Class Panel Graphs* had been successfully used to present the results to the experts to support a class-conceptualization process, and to assess the profiles associated to the classes.

In the field of multivariate analysis, Principal Component Analysis (PCA) is used to find a reduced set of principal components catching the relevant information contained in the data set. In this field it is common to interpret the principal components to elicit latent attributes implicitly measured in the data set. Contribution of original attribute to the principal components is used for this purpose and this is measured by using the angles of the attributes projected over the factorial space, which are directly linked to the correlation between attributes and principal components. Additionally, [7] introduce a statistical test (*Test-Value*) that provides objective criteria to identify the main contributing attributes to a certain principal component. They are based on the comparison of

means/percentages within the classes with respect to the global sample indicator. Although *Test-Value* comes from the multivariate analysis field, we think it can help in the interpretation of clusters as it will be shown along this work.

In this field, we have not found many research works in the literature. Methods or tools to obtain or help the cluster interpretation seems to be a problem not yet solved.

2. Methodology

The main idea is to find a connection or concordance between the interpretation that a technician/expert makes by reading *Class Panel Graphs* and a set of statistical tests expected to show significances in the same attributes and classes as experts used in interpretation.

After a clustering process, the result is a partition that defines the different clusters or *classes* of the individuals. This resulting partition can be used as a new categorical attribute and therefore, this attribute can be compared against other attributes using statistical tests. Among all attributes, the *active* attributes are those used to construct the clusters and the *illustrative* ones are those who have not been used to the construction of these clusters but can participate in the interpretation process to enrich cluster description. In this work, the following tests are assessed to relate them with the experts based interpretation obtained from direct reading of the *Class Panel Graphs*:

1. Kruskal-Wallis test for both active and illustrative numerical attributes. Given a numerical attribute X and the partition C, the statistics for the test is given by:

$$(n-1)\frac{\sum_{c=1}^{P} n_c(\bar{r}_c-\bar{r})^2}{\sum_{c=1}^{P}\sum_{j=1}^{n_c}(r_{cj}-\bar{r})^2} \sim \chi^2_{n-1} \quad (1)$$

where, n is the total number of observations, P is the number of groups of C, n_c is the number of observations in group c, r_{cj} is the rank of observations j from group c. $\bar{r}_c = \frac{\sum_{j=1}^{n_c} r_{cj}}{n_c}$ is the mean of the ranks assigned to a group c and $\bar{r} = \frac{1}{2}(n+1)$ is the average of all the r_{cj}.

Decision rule: Given a risk $\alpha \in [0,1]$, $p\_value < \alpha \rightarrow X$ and C have some association.

2. χ^2-Independence test for both active and illustrative qualitative attributes. Given two categorical attributes (X & C), the statistic for the test is given by:

$$\sum_{x=1}^{C_X}\sum_{c=1}^{C_C} \frac{(n_{xc}-\frac{n_x n_c}{n})^2}{\frac{n_x n_c}{n}} \sim \chi^2_{(C_X-1)(C_C-1)} \quad (2)$$

where, C_X is the number of categories of X, C_C is the number of categories of C, n is the number of observations, n_x is the number of observations with the xth category of X, n_c is the number of observations with the cth category of C and n_{xc} is the observed number of observations with the xth category of X and the cth category of C.

Decision rule: Given a risk $\alpha \in [0,1]$, $p\_value < \alpha \rightarrow X$ and C have some association.

Generally, the non-significant attributes would not be required for class description. However, in the application it will be seen how in some cases Kruskal-Wallis and χ^2-Independence test are not sensitive enough, sometimes for a high asymmetry, or too small classes or other causes. For this reason, a second test is performed: *Test-Value*, which is a new strategy in cluster interpretation process.

3. *Test-Value* for those attributes non-significant in previous tests. *Test-Value* finds the relevant numerical attributes for each class by comparing the global mean with the mean

within the class. Similarly, for qualitative attributes, compares the global proportion of one category with the proportion of this category within the class. Thus, this test allows finding relevant attributes per each class. To characterize a class with numeric attributes X, *Test-Value* is computed by the following statistics:

$$\frac{\bar{x_c} - \bar{X}}{\sqrt{(1 - \frac{n_c}{n})\frac{s^2}{n_c}}} \sim t_{n_c - 1} \qquad (3)$$

where, $\bar{X}$ is the mean of attribute X, $\bar{x_c}$ is the mean of X within the class c, n is the number of objects in the total sample, n_c is the number of objects in class c and s is the standard deviation of X.

Similarly, to characterize a category of a qualitative attribute the following statistic is computed:

$$\frac{\frac{n_{sc}}{n_c} - \frac{n_s}{n}}{\sqrt{(1 - \frac{n_c}{n})\frac{\frac{n_s}{n}(1 - \frac{n_s}{n})}{n_c}}} \sim Z \qquad (4)$$

where, n is the number of objects in the sample, n_s is the number of objects in the sample with the category s, n_c is the number of objects of class c and n_{sc} is the number of objects in class c with the category s.

In addition, these tests are solved using critical points ($critical point < |value|$). Thus, if a significant value is negative then the mean or proportion of the class is lower than the global one, and if it is positive, it is higher.

This *Test-Value* test has been finally used for all attributes because it is a good indicator of how the classes behave in the analyzed attribute whereas χ^2-Independent test and Kruskal-Wallis test only assess the global significance of an attribute without giving precise information on which change(s) is the one having different values (that can be higher or lower than the other classes). The purpose of using this test is twofold: first, to address the lack of sensitivity that can have the previous tests for non-significant attributes and second, to add information about the behaviour of the significant attributes.

Summarizing, *Test-Value* is used to see how the attributes distinguish every single class from the general trend of the attribute. As previously indicated, this test comes from multivariate statistical techniques and we introduce it in the clustering interpretation process for the first time in this work.

3. Experiments and Results

The data used in this work come from a real sample of 90 persons from general population in Catalonia participating in a nutritional study. Among other information, habits have been registered through a diet and physical activity questionnaire.

The subject profiles depending on their diet and physical activity habits before the nutritional study are searched.

The diet habits are described through 14 categorical attributes; these attributes have two categories to answer if the person consumes more than certain quantity of certain food “per day” or “per week” depending on the type of food.

The physical activity is described through 10 numerical attributes; from these attributes, 4 attributes were included in the analysis because the rest contains redundant information. The level of physical activity is asked as the calories spent ”per week”.

A clustering process was applied over these 18 attributes of diet and physical activity habits to identify subjects profiles. The data set contains a mixture of categorical and numerical attributes.A hierarchical clustering with the Wards' Method [8] and Gower's metric [9] have been used to build the clusters. The resulting dendrogram was cut in 3 classes, according to the Calinsky-Harabasz index [10]. This partition has been interpreted and validated by technicians/experts by using the *Class Panel Graphs* [11] as a support interpretation tool. Tables 6 & 7 show the resulting *Class Panel Graphs* of the data against the resulting classes and the resulting interpretation is given below.

Reference experts' interpretation:

H_1 is a group that eats more vegetables but less fresh fruit. Comparing with other groups, they consume a lower quantity of red meat, less legumes and commercial bakery. In addition, this group 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 groups, 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 groups. This group is the one which practices less physical activity.

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

In the following, the performance of several statistical tests is assessed to see how they can recognize the relevant class-characteristics observed by the experts over the *Class Panel Graphs*.

Table 1 shows the results of the Kruskal-Wallis test (KW) for numerical attributes and χ^2-Independence test (χ^2) for qualitative attributes against the classes. Given a risk level α, the attributes are considered significant when $p\_value < \alpha$. In this work $\alpha = 0.1$ has been used. Significant attributes are marked with (*) in the table.

Table 1. *p*-values of attributes vs classes

Attribute	Type	Test	*p*-value	Attribute	Type	Test	*p*-value
mainOliveOil	Q	χ^2	0.01768*	whiteMeat	Q	χ^2	0.3712
oliveOil	Q	χ^2	0.2315	sauce	Q	χ^2	0.5165
vegetables	Q	χ^2	6.572e-06*	lightWeek	N	KW	0.3625
fruit	Q	χ^2	1.97e-07*	moderateWeek	N	KW	0.02238*
redMeat	Q	χ^2	0.0001269*	intenseWeek	N	KW	0.4786
butter	Q	χ^2	0.06948*	homeWorkWeek	N	KW	0.992
gasDrinks	Q	χ^2	0.1146	totalWeek	N	KW	0.009571*
wine	Q	χ^2	0.3319	lightYear	N	KW	0.3216
legume	Q	χ^2	0.2164	moderateYear	N	KW	0.02464*
fish	Q	χ^2	0.0001266*	intenseYear	N	KW	0.2701
commercialBakery	Q	χ^2	3.248e-06*	homeWorkYear	N	KW	0.9958
nuts	Q	χ^2	0.0001091*	totalYear	N	KW	0.01987*

N: numerical Q: qualitative *: $p - value < 0.1$

These tests can identify the attributes that register important changes among classes. Significance means that the attributes have a notable different behaviour in at least one class. Table 3 shows the degree of concordance between attributes used by experts in their description against significance assessed by statistical test and it can be found a

good approach. However, the drawback of these tests is that given a significant attribute, the test result itself does not provide details on which class/es is/are behaving differently than the others. For instance, the attribute *fruit* is selected as significant, but the test do not gives indications on which classes contain people consuming more or less fruit. Multiple comparison techniques are required to assess these issues. This is an expensive method because a combinatorial number of tests must be performed for every significant attribute.

As said in section 2, the *Test-Value* is imported from the multivariate analysis field to go one step further in the interpretation of classes. For each attribute the *Test-Value* of every class is shown in Table **??**. It can be seen that most of the attributes which are not significant for general tests (Kruskal-Wallis and χ^2-Independent test) are also not significant. Only two cases show significance with *Test-Value* which were not detected as significant by the general tests: first, the attribute *gasDrinks* is significant for class H_2 and secondly, *legume* is significant for the class H_1.

Table 2. *Test-Value for both Active and Illustrative Attributes*

	Significant?			*Test-Value*			*Critical Point*		
Attribute	Type	KW/χ^2	Condition	t_{H_1}	t_{H_2}	t_{H_3}	H_1	H_2	H_3
mainOliveOil	Q	Y	yes	1.383	-2.841*	0.969	1.645	1.645	1.645
oliveOil	Q	N	≥4spoon	-0.301	-1.325	1.469	1.645	1.645	1.645
vegetables	Q	Y	≥2day	4.777*	-3.22*	-2.355*	1.645	1.645	1.645
fruit	Q	Y	≥3day	-5.025*	0.375	5.084*	1.645	1.645	1.645
redMeat	Q	Y	≥1day	-1.424	4.178*	-2.082*	1.645	1.645	1.645
butter	Q	Y	≥1day	0.432	1.771*	-1.997*	1.645	1.645	1.645
gasDrinks	Q	N	≥1day	-0.539	2.018*	-1.166	1.645	1.645	1.645
wine	Q	N	≥7week	-0.368	-1.078	1.328	1.645	1.645	1.645
legume	Q	N	≥3week	-1.734*	1.048	0.959	1.645	1.645	1.645
fish	Q	Y	≥3week	2.763*	-4.15*	0.617	1.645	1.645	1.645
commercialBakery	Q	Y	≥2week	-4.462*	0.149	4.673*	1.645	1.645	1.645
nuts	Q	Y	≥3week	-0.549	-3.433*	3.56*	1.645	1.645	1.645
whiteMeat	Q	N	yes	-1.093	-0.243	1.387	1.645	1.645	1.645
sauce	Q	N	≥2week	0.383	-1.133	0.568	1.645	1.645	1.645
lightWeek	N	N		0.713	-1.278	0.338	1.682	1.74	1.703
moderateWeek	N	Y		-1.28	-1.618	2.777*	1.682	1.74	1.703
intenseWeek	N	N		0.792	-1.308	0.28	1.682	1.74	1.703
homeWorkWeek	N	N		-0.667	0.558	0.235	1.682	1.74	1.703
totalWeek	N	Y		0.107	-2.627*	2.157*	1.682	1.74	1.703
lightYear	N	N		-0.001	-0.794	0.687	1.682	1.74	1.703
moderateYear	N	Y		-0.721	-1.823*	2.353*	1.682	1.74	1.703
intenseYear	N	N		0.281	-1.373	0.885	1.682	1.74	1.703
homeWorkYear	N	N		-0.569	0.604	0.089	1.682	1.74	1.703
totalYear	N	Y		-0.197	-2.298*	2.201*	1.682	1.74	1.703

Expert' description of class H_1 contains references to attributes: *vegetables, fresh fruit, red meat, legumes and commercial bakery*. From those, all except *legumes* where identified by general χ^2- Independence test, but *Legumes* is also retrieved when *Test-Value* is used. In fact, the *Class Panel Graph* shows a smaller proportion of people eating frequent *legumes* than the other classes and experts agree that this attribute has to be included in the description. Similarly, the Class Panel Graph show a higher proportion of frequent *gasDrinks* intake for class H_2, which was not identified by general χ^2-Independence test but appears as significant with *Test-Value*.

Table 4 shows better degree of concordance between expert's criteria.

Table 5 shows a summary of the significant attributes for both global test and *Test-Value*. In column "KW/χ^2", significance is marked with "*". Significance in *Test-Value* is marked with "+" when class shows significantly higher values than the average or with "-" when it shows significantly lower values than the average. From this table 5, it is

possible to observe that all significant attributes for Kruskal-Wallis or χ^2-Independence tests are also significant for at least one *Test-Value*. Thus, it seems that *Test-Value* tests are more expressive than the global tests and in no-cases the global test is significant when the *Test-Value* are non-significant. For this reason we intend to test in future application if this is a general property and *Test-Value* can be directly used for interpretation and global tests can be directly skipped.

Table 3. Experts' Description vs General Tests

		Kruskal-Wallis & χ^2-Indep. tests	
		Yes	No
Experts' Description	Yes	main Olive Oil, Vegetables, Fresh Fruit, Red Meat, Butter, Fish, Commercial Bakery, Nuts	gas Drinks, Legumes
	No		Olive Oil, Wine, White Meat, Sauce

Table 4. Experts' Description vs *Test-Value*

		Test-Value	
		Yes	No
Experts' Description	Yes	main Olive Oil, Vegetables, Fresh Fruit, Red Meat, Butter, Legumes, Gas Drinks, Fish, Commercial Bakery, Nuts	
	No		Olive Oil, Wine, White Meat, Sauce

From the whole description provided by the experts only in one case *Test-Value* are not sensitive enough, from the *Class Panel Graph* H_1 is consuming much less *redMeat* than H_2 but the test is not detecting this fact. Further analysis is required to clarify if these cases can also be detected with other testing procedures.

Table 5. *Summary of the Attributes Significance*

		Test-Value					*Test-Value*		
Attribute	KW/χ^2	t_{H_1}	t_{H_2}	t_{H_3}	Attribute	KW/χ^2	t_{H_1}	t_{H_2}	t_{H_3}
mainOliveOil	*		-		whiteMeat				
oliveOil					sauce				
vegetables	*	+	-	-	lightWeek				
fruit	*	-		+	moderateWeek	*			+
redMeat	*		+	-	intenseWeek				
butter	*		+	-	homeWorkWeek				
gasDrinks			+		totalWeek	*		-	+
wine					lightYear				
legume		-			moderateYear	*		-	+
fish	*	+	-		intenseYear				
commercialBakery	*	-		+	homeWorkYear				
nuts	*		-	+	totalYear	*		-	+

4. Conclusions and Future Work

In this work clustering techniques have been used to find profiles in dietary and physical activity habits of a real sample from a general population in Catalonia. *Class Panel Graphs* have been used to interpret the profiles emerging from clustering. With the aim of contributing to a more efficient interpretation process that can be objectified and automated in the future.

However from the whole description provided by the experts only in one case *Test-Value* are not sensitive enough. From the *Class Panel Graph*, H_1 is consuming much less *redMeat* than H_2 but the test is not detecting this fact. Further analysis is required to clarify if these cases can also be detected with other testing procedures.

Statistical tests are introduced as a post-processing of clustering results to retrieve from clusters the relevant attributes found by the experts.

We have tested how the statistical tests can assist or help the interpretation of the classes. As a first step classical global tests like χ^2-Independence test and Kruskal-Wallis test were used to identify relevant attributes for the clustering. However, some attributes retained as important by experts are not assessed as significant by those global tests. *Test-Value* has been imported from PCA field and used in the context of clustering, and it seems to be more sensitive than global tests by approaching better the experts' interpretation. They are also more expressive and give more precise information on which attribute behaves differently in which class with clear indication about the sense if this difference shows higher or lower values than average.

However, it seems that *Test-Value* is more expressive than the global tests and in no-cases the global tests are significant when the *Test-Value* is non-significant. For this reason, we intend to test in future application if this is a general property and *Test-Value* can be directly used for interpretation and global tests can be directly skipped.

Acknowledgements

Beatriz Sevilla thanks to the ACIA fellowship for supporting her attendance to CCIA'2013.

References

[1] K. Gibert and Z. Sonicki, "Classification based on rules and medical research," *Journal of Applied Stochastic Models and Data Analysis, formerly JAMSDA*, vol. 15, no. 3, pp. 319–24, 1999.

[2] M. Sarstedt and E. Mooi, *A Concise Guide to Market research: The process, data, and methods using IBM SPSS statistics*. Springer Verlag, 2011.

[3] W. Cecere and D. A. Abreu, "A method for improving list building: Cluster profiling," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2010.

[4] D. Haughton, P. Legrand, and S. Woolford, "Review of three latent class cluster analysis packages: Latent gold, polca, and mclust," *American Statistician*, vol. 63, no. 1, pp. 81–91, 2009.

[5] G. R.-S. Karina Gibert, Alejandro Garca-Rudolph, "The role of kdd support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation," *Acta Inform Med*, 2008.

[6] K. Gibert, C. Garca-Alonso, and L. Salvador-Carulla, "Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support," *Health research policy and systems BioMed Central*, vol. 8, no. 28, p. 28, 2010.

[7] L. Lebart, M. Piron, and A. Morineau, *Statistique exploratoire multidimensionnelle*. Dunob, 3 ed., 2000.

[8] C. de Rham, "La classification hirarchique ascendante selon la mthode des voisins rciproques," *Cahiers de l'analyse des donnes*, vol. 5, no. 2, pp. 135–144, 1980.

[9] J. Gower, "A General coefficient if similarity and some of its properties.," *Biometrics*, vol. 27, pp. 857–874, 1971.

[10] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, pp. 1–27, 1974.

[11] B. Sevilla Villanueva, K. Gibert, and M. Sànchez-Marrè, "Clustering and interpretation on real nutritional data," in *Conference VII Simposio de Teoría y Aplicaciones de Minería de Datos*, 2013.

Table 6.: Diet characteristics vs Classes

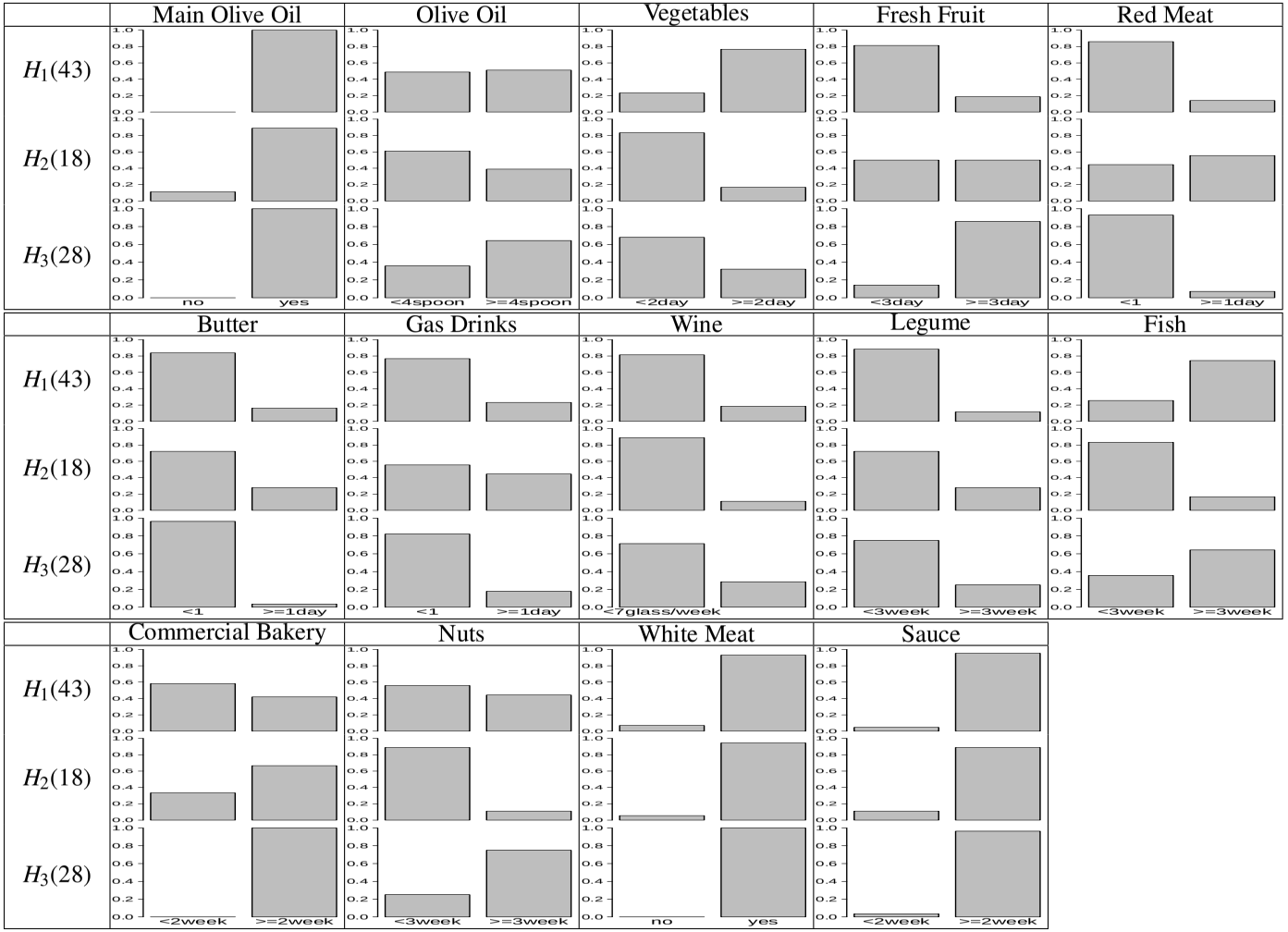

Table 7.: Physical Exercise characteristics vs Classes