The K-modes algorithm applied to Gender Analysis

Author: Neus Llop Torrent.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Sònia Estradé Albiol, Enric Pérez Canals.

Abstract: The aim of the project is to find an objective tool to analyze relevant data on gender studies. Over the total proven methods, K-modes has been chosen as the most suitable tool for the analysis due to the categorical nature of the variables in the data-set. The algorithm implementation produces sub-groupings of the samples depending on the dis/similarities between them. The input data were obtained from a survey answered by Spanish physics professors.

I. INTRODUCTION

Whereas the general tendency over the science faculties in Spain shows an increase of the female professors percentage, numbers still show low participation in physics faculties. In 2013 at the University of Barcelona there was 44,9% of women as PDIs while in 2018 the number increased up to 46%.

Nevertheless, examining all departments of the University of Barcelona, the last four positions regarding women jobs percentages come from the Faculty of Physics. Furthermore, the Faculty of Physics shows the second inferior percentage of women as professors with 18% [1]. Comparing the figures with the overall physics universities in Spain the percentage is pretty similar. Besides, research carried out by the Spanish Ministry of Science reflected that in scientific institutions such as the Spanish National Research Council (CSIC), during the same period, the women percentage maintained at 42 %. However, the grade A workplaces were stagnant in man with 75%[2].

	UB
Applied Physics	24
Quantum Physics and Astrophysics	19
Condensed Matter Physics	19
Electronic Engineering	15

TABLE I: Comparative table of % of women working in physics departments at the University of Barcelona (UB).[1][2]

In response to these relevant statistics, the Gender Equality Association of the Faculty of Physics from the University of Barcelona designed a survey to go beyond the classic gender analysis while looking for hidden patterns among the community of professors. The aim of the study was to distribute the surveyed sample in different groups based on the dis/similarity amongst them, depending on the responses they had provided. The dimension of the data, which corresponds to the number of questions, required computational techniques due to the time and volume of the operations involved. For all the reasons listed above, the machine learning clustering analysis is proposed [3]. The most challenging issue to settle is the choice of the right algorithm to deal with the categorical nature of the data.

II. PATTERN RECOGNITION METHODS

Two approaches could be followed assuming that the objective of the project is to find hidden patterns from a multiple-choice questionnaire. Both statistics and machine learning methods share the objective of learning from the data. Choosing a machine learning method rather than the classic statistics would be justified because, unlike the second one, the first one does not require previous suppositions on the correlations between variables [4]. Moreover, this is expected to be a forward-thinking project that aspires to seek the possible benefits provided by machine learning towards gender analysis.

A. Choosing the right algorithm

One of the challenges that machine learning involves is choosing the right algorithm. In the first stage, it is divided into two learning methods, supervised and unsupervised. While supervised learning requires knowing the answer to some of the data so that the algorithm could be trained, in unsupervised learning the output is unknown [4]. Therefore, unsupervised learning will be more appropriate due to the small amount of previous information. Commonly used to group unlabelled data, clustering unsupervised learning methods will be considered from now on [3].

B. From k-means to k-modes

In order to obtain the right clustering algorithm, it is important to take into account the number of output clusters expected, its size, and geometry. Accordingly, K-means clustering is proposed, due to the low number of clusters we are expecting from the analysis and the not very large data-set.

Furthermore, when there is a belief in the variation in the behaviour of different subgroups, K-means is used. The method is particularly useful when the data has so many dimensions that it is not possible to visualize the dependencies in a graphic.

1. K-means

K-means algorithm divides m points within n dimensions into k clusters, so that the within-cluster sum of squares is minimized. That means that the process can be understood as an optimization problem, where the Euclidean distance is the objective function to be minimized. From now on, points will be represented as vectors (1) in the n dimension space [5].

$$(x_1, x_2, x_3, ..., x_n) \tag{1}$$

The algorithm works through the steps listed below:

1. The algorithm starts by picking up some random points that will be considered the first cluster centers, also known as centroids (2). It is important to remark that K-means random starting points do not have to be in the data-set as long as they are in the data space.

$$(c_1, c_2, c_3, ..., c_k) = centroids \tag{2}$$

2. Once centroids have been selected, K-means will build k different groups. To achieve that goal, the algorithm iterates over all points and calculates the Euclidean distance (3) between it and all the centroids to finally assign it to the nearest one. Euclidean distance is defined as absolute value of squared difference between each point (1) to its centroid (2).

$$D(x_i, c_i) = \sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \tag{3}$$

To quantify how confident we are in the cluster assignments rather than simply allocate every point to the closest cluster, responsibility (4) is calculated. Responsibility measures the probability of every point to pertain to a certain cluster.

$$r_k^n = \frac{exp(-\beta D(c_k, x^n))}{\sum_i exp(-\beta D(c_k, x^n))} \tag{4}$$

3. When all points have been assigned to a cluster, centroids will be recalculated considering the points in that moment belonging to the cluster. In order to know which is the new center of the cluster, the mean (5) of all points from the same cluster is computed. The mean of a set of vectors is easily obtained by adding them up and dividing by the number of vectors. However, every data point would be influenced by the responsibility (4) and as a consequence regular mean transforms to weighted arithmetic mean. As a result, a new centroid is obtained.

$$c_k = \frac{\sum_n r_k^n x^n}{\sum_n r_k^n} \tag{5}$$

where $r\epsilon[0, 1]$

4. Once all centroids have been calculated we go back to step 2 and repeat the two steps mentioned before over and over until they converge to an answer. In order to know when an optimal solution has been reached, an objective function (6) is created. Defined as the sum over the n dimensions of the sum over the k clusters of the Euclidean distance (3) weighed with the matching responsibility (4), it is to be considered the key in the points reassignment. That could be also understood as an optimization problem where the minimized function will define the end of the loop [6].

$$\phi = \sum_n \sum_k r_n^k ||c_k - x^n||^2 \tag{6}$$

Over the iterations, it is mathematically guaranteed that the objective function (6) will always decrease to a local minimum [5].

While the theory is clear, the biggest problem appears in relation to our data-set: K-means algorithm is not directly applicable to categorical data. Consequently, Euclidean distance in which K-means is based does not make sense in a discrete space. To overcome the problem two methodologies can be followed.

From the k-means basis, the first method consists in making a pre-processing to transform the categorical values into binary.

Owing to the high execution time and computational memory required, this method was discarded. Therefore, an alternative method must be explored.

2. K-modes

The second approach proposed is an extended version of K-means, called K-modes. K-modes algorithm works very similarly to K-means, though, instead of using the mean (5) to calculate the centroid it uses the mode [7]. In other words, the clusters will be defined based on the number of matching categories between data points that means using the highest frequency to form the clusters. As in more categories two points overlap, the higher their probability to belong to the same cluster [8]. Differences from both algorithms are commented subsequently.

1. Just like in k-means, the first step would consist of assigning randomly a number of points as the transitory k modes.

2. During the points clustering assignment, while in K-means the Euclidean distance (3), in K-modes

this step consists in calculating the dissimilarity (8) score between each of the remaining data points from the k number of clusters chosen [9]. The criteria to say that two points differ or not in the i variable (7) is as follows:

$$\delta(x_i, c_i) = \begin{cases} 0 & \text{if} \quad x_i = c_i \\ 1 & \text{if} \quad x_i \neq c_i \end{cases} \tag{7}$$

Dissimilarity (8), also known as the Hamming distance, is defined to extend the comparison to every pair of variables [3]. For instance, if two points have the same attributes their dissimilarity will be 0 whereas if they do not share any attribute dissimilarity will be n [9].

$$d_{xc} = \sum_{i=1}^{n} \delta(x_i, c_i) \tag{8}$$

where x_i refers to a data-set vector and c_i to a centroid.

3. Unlike the K-means algorithm, where the centroid re-assignment was computed through the mean, in K-modes it is replaced by the mode. Mode, defined as the most frequent value in the cluster, will work as the new cluster center [7]. Therefore, during iteration, each point will be associated with the mode whose score is minimum.

   After points assignment, it is necessary to recalculate the mode for each cluster using the frequency method (7) to update the centroid

4. Finally, as in K-means, we iterate over the last two steps until there is no new cluster re-assignment.

C. Cost Function

Both K-means and K-modes drawback is that they need as an input the final number of clusters in which we want to split the points. As we are dealing with unsupervised learning and there is no previous information, a priori we do not know the best suitable number of subgroups. The process to find out the optimum number is called the Elbow Method and works trough computing the cost function (9).

For K-means, the cost (9) is defined as the sum of squares error within the cluster and gives information on how scattered the points from a cluster are. Therefore, the lower the cost, the nearer the points in the cluster [10]. Nevertheless, to compute the cost for K-modes the Euclidean distance (3) has to be replaced for the Hamming distance (8).

$$cost = \sum_{i=1}^{n} \sum_{i=1}^{k} d_{xc} \tag{9}$$

By plotting the cost function against the number of clusters an elbow should be found. During the clusters number growth, there is a point where the drop starts to change smoothly and the increase of k does not give significant improvements. The number where the cost (9) begins to slightly decrease is the number that best fits the data-set sub-grouping [10].

III. EXPERIMENT AND RESULTS

The experiment was generated through a multiple-choice survey answered by a total of 168 professors belonging to the University of Barcelona Faculty of Physics, the Institute of Cosmos Sciences (ICCUB) the Association of research and technologists women (AMIT) and the women group of the Royal Spanish Society of Physics(RSEF). The questionnaire was subdivided into 14 questions with a variable number of responses.

For data processing, the implementation of the algorithm and the representation of the results was carried out using the Python SciKit Learn library .

A. Pre-processing

In order to refer to the data after the processing a more appropriated technical vocabulary is used. Every individual in the data-set is represented in the matrix as a row whereas every question is understood as a feature with its corresponding column associated. Notice that every question (feature) is represented with different responses, called attributes.

Due to the fact that the raw material is text questionnaire and the answers are in a categorical format, a pre-processing is needed.

Nevertheless, K-modes do not require the intermediate step of transforming into vectors the questionnaire attributes. Then, the input data must be convert to another format so that the operations that will be applied to the data-set make sense. To transform every response into a data point the label encoder function convert the text categorical features into a numerical representation. Finally, a numerical matrix (168∗14) is obtained.

B. Testing survey

To test the algorithm implementation, 48 professors from the University of Barcelona answered a survey with reduced number of 11 questions. Owing to the low number of samples the results could not be completely accurate, however, it helped in the process of preparing a more suitable survey. There were two main problems: First, some physicists were confused by the way some questions were asked resulting in homogeneous answers. As a consequence, some features ended up having almost a unique attribute. Second, a significant number of questions were left unanswered.

The algorithm interprets that blank questions as coinciding amongst them but different from the attributes given before in the computed dissimilarity (8). When the matrix has missing values they will be understood as NaN and a new category is created.

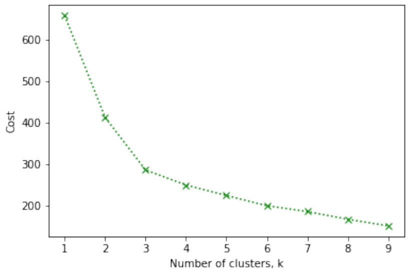

FIG. 1: Elbow plot. 11 dimensions and 48 samples. Considering all individuals.

Consequently, if a high percentage of people give missing answers, that will result in a miss-classification as one category will consist mainly of the people with the NaN attributes. To solve this problem in the first analysis the one-category features were eliminated.

C. Final survey and Results

With the information obtained from the first experiment, a second survey was designed to deal with both problems. First, the corrections included a rewriting to make questions more understandable. Second, it was required that the online survey could not be sent if there was any missing answer. Furthermore, three questions related to parenthood were added to see which different labeling of the data-set would produce.

Surprisingly, while the optimum clusters number for the test survey was three (FIG. 1), four labels were found in the final one (FIG. 2). The Elbow-Method was performed through three different initialization routines. Methods differ in the selection of the points that will be considered the first centroids. Huang [7] uses the frequency method (8) explained in the previous section while the Random method selects an arbitrary set of points from the data-set. However, those processes do not consider the attributes relative frequencies in the cluster centroid, producing groups with more disparate objects [8]. The final results are based on the Cao method which differs in considering the density of an attribute in a cluster. The density is estimated through the distance from that attribute to the other ones in the data-set [11].

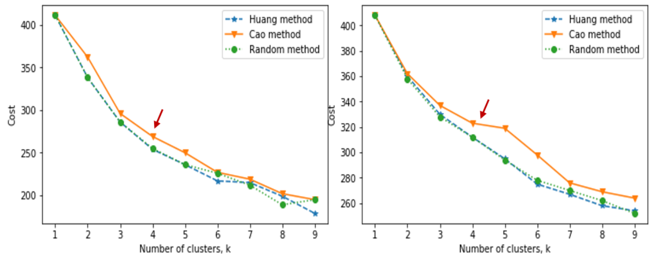

FIG. 2: Elbow plot. Left: 9 dimensions and 169 samples. Individuals with partner. Right: 12 dimensions and 120 samples. Individuals with a partner and children.

For the algorithm implementations (FIG. 2), 9 dimensions, 169 samples, and four output clusters were considered. The algorithm was also executed without the gender and partner rows to prevent the algorithm to group people by gender. Still, when results are presented both features are recovered.

Results are separately represented with information of the most relevant contributions. The high number of dimensions makes it difficult to plot all dependencies. For this reason just the relevant discrepancies between groups are exposed. In (FIG. 3), a two dimensions basic study is displayed regarding the companions studies of the individuals on a relationship. A second three dimensions chart on (FIG. 4) analyses, for the same individuals, their partners workplace and the gender whom they feel more comfortable working with. It can be easily observed from (FIG. 2) that the appropriate clustering was unclear with the smallest data-set. Moreover, in the Elbow plot when the number of the dimensions increases the labeling quality worsens too. Accordingly, the best data interpretation was done through the longest data-set and with a reduction of the dimensions. Results from

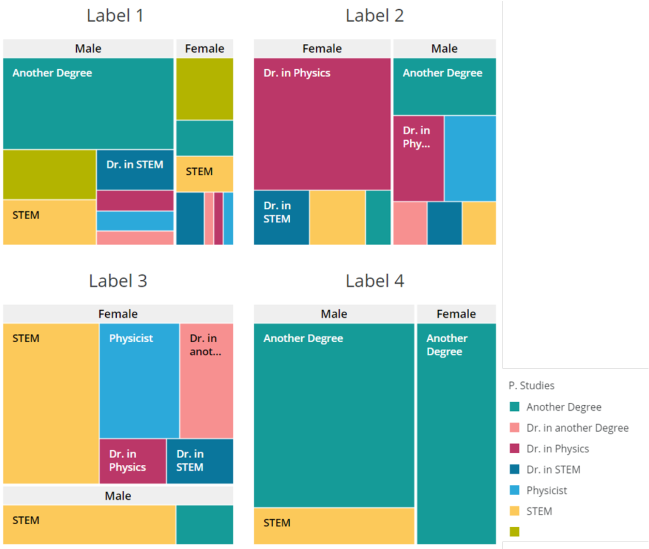

FIG. 3: The heat map displays information on the job distributions by gender on every cluster. The size of the boxes is directly proportional to the percentages of individuals with partners with these studies.

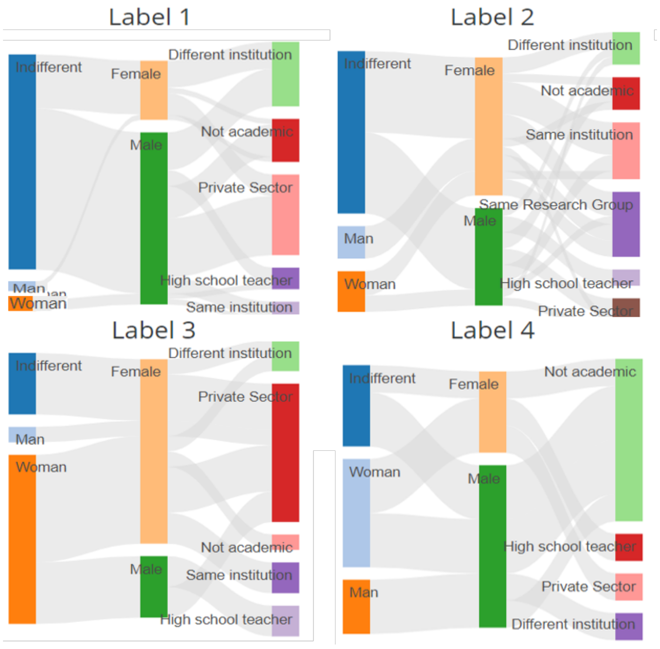

FIG. 4: The illustration shows a Sankey diagram which represents the flow between three features from the survey by label. To the left, the centered gender distribution flows to the surveyed opinion about their gender preferences when working with a research team. To the right, the gender distribution flows to the workplace of their partners. The widths of the bands are directly proportional to the percentages.

(FIG. 4) point towards a high likelihood that one gender is dominant in each group. It should be noted that the algorithm did not have information on the surveyed gender or partner gender which implies that results could not be anticipated. As a consequence, it is inevitable to appreciate that the patterns on the answers let the conclusions become gender-based differences analysis. Despite this fact, it can still be stated that individuals from equal gender have been distributed into two groups. Considering labels with a male preponderance (1 and 4), (FIG. 3) revealed that in both, their partner study was mostly not Science neither STEM-related. Looking now to the female gender (Labels 2 and 3), most of the partners are shown to have scientific or technological education (Physics or STEM). Another pattern to consider is from a meaningful group of women working in their partners research group/ institution.This group seems to do not have a preference for gender regarding their workplace team, whereas women with their partners working in a different institution or in a private sector have a preference with the female gender. On a similar direction, striking results have been found considering whether the surveyed physicists considered female-majority working groups being more collaborative and integrative. It has been found that women working in their partner research group were against it while women with partners elsewhere (label 3) agreed. Contrarily, results comparing more dimensions by considering children did not seem to show any significant patterns. Only, as expected women research time was more harmed by parenthood than, generally, it was for men.

IV. CONCLUSIONS

The clustering approach used for data processing has been especially valuable to encounter the patterns that involved more than two dimensions. However, for shorter data sets, it did not contribute to go beyond the expected results. Nevertheless, the clustering perspective can be regarded as being less subjective through not making preferences on the election of features that provide the final result. Another point that supports this research is the possibility to make all the matches without contemplating gender a priori. On the other hand, to achieve more accurate results, a new questionnaire may be proposed to deal with one-category features, explained because of the frequency based method. Hence, the more heterogeneity in the answers, the better the match of attributes. As a consequence, this survey should be planned to avoid questions with high probability to find a uniform response.

Acknowledgments

I would like to thank my tutors, Sònia Estradé Albiol, and Enric Pérez Canals for their support and guidance on the project.

[1] G. tècnic Rectorat, *Memòria del curs 2017-2018* (Edicions de la Universitat de Barcelona, Barcelona, 2019).
[2] A. P. Rodríguez, *Científicas en cifras 2017* (Ministerio de Ciencia, Innovación y Universidades, Madrid, 2018).
[3] K. Jain et al., ACM Comput. Surv. **31**, 264 (1999).
[4] R. Golden, *Statistical Pattern Recognition* (Pergamon, Oxford, 2001).
[5] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, vol. 1 (University of California Press, Berkeley, Calif., 1967).
[6] Kanungo et al., IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 881 (2002).
[7] J. Z. Huang, *Clustering Categorical Data with k-Modes* (Information Science Reference, 2009).
[8] J. Z. Huang, Data Min. Knowl. Discov. **2**, 283 (1998).
[9] Z. He et al., *Approximation Algorithms for K-modes Clustering*, ICIC' 06 (Springer-Verlag, Berlin, Heidelberg, 2006).
[10] D. J. Ketchen and C. L. Shook, Strategic Management Journal **17**, 441 (1996).
[11] F. Cao, J. Liang, and L. Bai, Expert Systems with Applications **36**, 10223 (2009).