A Biased Random-key Genetic Algorithm with Local Search Applied to Unsupervised Clustering of Cultural Data **Applications Track**

A. H. Aono[1], R. M. de Oliveira[2], B. O. Franchi[1], J. S. Nagai[1], H. E. S. Paz[1], A. A. Chaves[1], C. B. Martins[1]

[1] Universidade Federal de São Paulo, Brazil
alexandre.aono@unifesp.br
[2] Universidade do Estado de Minas Gerais, Brazil

Abstract. The relationship between culture and development has come to occupy a prominent place in the present time. From this perspective, culture is commonly said to be a constructive axis of identities, and also an economic factor which generates wealth. Due to the ongoing need of implementing an institutional policy for culture at Federal University of São Paulo, it was created a culture and extension project with the objective to contribute with the cultural actions in the college campuses. Sociocultural data from the university students was collected and herein we present a different methodology of analyzing such kind of dataset. Using the metaheuristic optimization Biased Random-key Genetic Algorithm with local search, we have achieved positive results in the identification of cultural subunits in the university. Significant information to understand the students' integration and different cultural habits was obtained and a new way of visualizing this cultural scenario is herein proposed.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: metaheuristic, optimization, multivariate analyses, cultural profile

1. INTRODUCTION

According to Merriam-Webster dictionary [Merriam-Webster Dictionary 2002], culture means '*the customary beliefs, social forms, and material traits or a racial, religious, or social group; also the characteristic features of everyday existence (such as diversions or a way of life) shared by people in a place or time*'. Culture also can be considered the set of artificial social ideas, behaviors and practices learned from generation to generation through life in society [Kroeber 1949] and education [Morgado 2016]. Through education individuals are exposed to new abilities and knowledge such as techniques, different ways of living, i.e., the culture of the group [Morgado 2016]. In such context, culture is one of the fundamental instruments in the formation of more tolerant, generous, sensitive and creative young people. Its acquisition and perpetuation is a social process, resulting from learning [Morgado 2016]. Thus it can be said that culture is an inseparable part of the formation of the individual. It is inclusion and a gateway for building a more understandable, tolerant, and humane society.

Using this concept of culture is important to demonstrate that particular modes of expression and social interaction find explanations in habits, customs and beliefs shared by members of the same group or society [Resende and de Paula 2011]. In addition, organizational culture can influence the construction of identity, since in the context of institutions these individuals are transformed as they adapt to the demands of the different social groups of which they are part. The relationship between

culture and development has come, increasingly and rapidly, to occupy a prominent place in the present time. From this perspective, it is important to add to culture's definition the constructive axis of identities, as a privileged space for achieving citizenship and social inclusion. [Canclini 2001] also corroborates the idea that culture can help solve social problems such as violence, unemployment, environmental degradation and social exclusion.

According to [Wang 2007], culture changes with changing economic, socio-political circumstances, commercial or political relations. However, cultures are constructed by people [Wang 2007]. Saying that the human being is a historical-cultural individual emphasizes the idea that the being is only human if inserted into a cultural group, which provides modes of human behavior. Thus, people are not mere objects of cultural influences, but subjects who can sift various influences and reject or integrate them [Wang 2007]. [Barth 1998] points that based on most anthropological reasoning rests, culture is discontinuous, i.e., there are aggregates of people who essentially share a common culture with interconnected differences. The integration of such groups, however, represents a key point on the functioning of the whole society. [Taylor 1991] defines structural integration as the presence of persons from different cultural groups in a simple organization and points the importance of looking beyond organization-wide profile data to proper understand it. Examining cultural mix by function, level, and individual work group represents a way to understand how groups are highly segregated.

In this scenario, in order to assist the Extension and Culture Chamber (CaEC) of Federal University of São Paulo (UNIFESP) in its actions, a project with the purpose of creating a Cultural Information Database was created and the identification of different cultural profiles would benefit the entire university by visualizing how the students are organized in the institution. Since culture is nothing but a way to describe human behavior, it would follow that there are discrete groups of people, i.e. ethnic units, to correspond to each culture [Barth 1998]. These groups might help the creation of specific cultural actions in order to enable the structural integration of these groups and the amplification of cultural activities trough students' habits.

2. PROPOSED APPLICATION

Since culture is learned through life in society and education [Kroeber 1949; Morgado 2016], the understanding of how different groups are organized in an educational institution and how they interact is the first step to comprehend how social, economic, cultural and educational characteristics might influence the integration of groups and society segregation. According to [Barth 1998], there are aggregates of people who share a common culture in the society, but there's also the presence of interconnected differences and therefore these differences are what separate those groups. Herein we are proposing the usage of a meta-heuristic optimization algorithm for unsupervised clustering in order to identify segregate groups using cultural datasets. Using the socio-cultural characteristics of a federal university students, the usage of such approach represents an innovation in anthropological studies which are considered mainly observational. This methodology might be used to corroborate or contrast systematic observations or theoretical explanations to the comprehension of social aspects.

As [Wang 2007] points, people are inserted into cultural groups, which provide modes of human behavior. In this paper, we are trying to identify these discrete groups in the university using their cultural characteristics. According [Barth 1998], the differences between cultures have been given much attention; however the constitution of ethnic groups, and the nature of the boundaries between them, have not been correspondingly investigated. In the paper's context, the characterization of students in different cultural groups might allow the academic community to understand which factors are most relevant to determine a social group, how cultural units are organized in the institution and what indicatives can be used to identify groups with different cultural characteristics.

3. BIASED RANDOM-KEY GENETIC ALGORITHM

The BRKGA was proposed by Gonçalves and Resende [Gonçalves and Resende 2011] and is a variation of the Random-Key Genetic Algorithms (RKGA) [Bean 1994]. The BRKGA represents a solution with a vector of random-keys, which are real numbers in the interval $[0, 1]$. This vector (also called chromosome) is not considered as a solution of the problem. Therefore, it is necessary to decode the vector to a real problem solution. For each problem, we define a specific decoder. This is a deterministic algorithm that takes chromosome information and returns a solution to the problem. The fitness of each solution is also computed by the decoder. The evolution process of the BRKGA is independent of the problem. A population of p random-key vectors evolves over a number of generations. In each generation, the population is sorted by the fitness. Then, a small group with the best p_e solutions in fitness values (elite group) are copied without modification to the population of the next generation. A number p_m of random-key vectors, randomly generated (mutants), are also introduced into this population. The remainder of the population ($p - p_e - p_m$ solutions) is produced through the process of crossover, by combining an elite parent with a non-elite parent of the current population. The parameterized uniform crossover [Spears and Jong 1995] is used in BRKGA.

The method randomly creates an initial population of random-keys vectors. Each vector has n random-key, such that, n is the number of objects in the data. The solution of the problem is accomplished by corresponding random-key values by decoder. In decoder, the interval $[0, 1]$ is divided by k groups and the clusters are created with the objects that have the random-key in this interval. After decoding the solution, the fitness is calculated through the objective function proposed in Babaki et al. [Babaki et al. 2014]. Generally, the objective function to be minimized is the within-cluster-sums-of-squares (WCSS). The WCSS calculate by the distance from the elements (x) of the group to its centroid (C), as in Equation 1.

$$Z = \sum d^2(x, C) \tag{1}$$

According [Babaki et al. 2014], the distance between all cluster points (x) to centroid (C) is equal the distance (d) between the elements of the same cluster, divided by the size of each cluster, as in Equation 2.

$$Z = \sum d^2(x, C) = \frac{\sum_{x_1, x_2 \in C} d^2(x_1, x_2)}{|C|} \tag{2}$$

Such that, $x_1 \neq x_2$ belongs to the cluster C, and $|C|$ is the cardinality of the cluster. Every pair of two points in C is included in the sum once, without repetition. The calculation of Z is done simply dividing a sum of distances of each cluster by the number of elements of this cluster. To accelerate the convergence process of the BRKGA, we implemented a local search heuristic, which is applied to all offspring generated by BRKGA. The local search is applied in the decoded solution. The solution found by the local search is not transferred to the vector of random-keys. This preserves the diversity of BRKGA. The local search shifts the object of a cluster to the others. Each object is placed in a different cluster. The nearby solution (S') receives the current solution (S) and accomplishes movements to the other clusters. When the neighbor solution (S') is better than the current solution, the current solution (S) receives the neighboring solution (S') and the best solution (S^*) is stored. This heuristic is run while the fitness of the solution is improved. Algorithm 1 presents the shift heuristic proposed in this paper.

Algorithm 1 Shift local search (S)

```
1: while (improve S ) do
2:     for (each object) do
3:         for (each cluster) do
4:             S' ← S;
5:             S' ← Move (an object to another clusters);
6:             if S' better than S then
7:                 S ← S';
8:                 STORE (the best solution S*);
9:         end for
10:    end for
11: end while
12: return (S*)
```

4. METHODS

In this session, we describe the data used in this study, its collection process, the project to which this study is linked, its respective approval in the ethics committee and how the BRKGA was used with these data types.

4.1 Extension Project

With the need to implement an institutional policy for culture at UNIFESP, an extension project was created with the purpose of tracing the cultural profile of the undergraduate students of the Institute of Science and Technology (ICT) of UNIFESP, assisting and guiding the cultural actions of the campus in order to diminish the cultural difference between the different cultural profiles of the campus. To this end, a database containing sociocultural information of undergraduate students has been created for further analysis. Besides, this project is part of an extension program entitled '+Cultura', which wants to promote activities in the entire campus of São José dos Campos and other public institutions in the region, supplying to students and the entire community means to promote culture. The objective of the project is collecting social and cultural data, statistically analysing the cultural profile of the campus, and creating a database platform to store this kind of data.

4.2 Dataset

The data used in the paper was collected using printed questionnaires. The project was cleared by the Institution's Ethics Committee (CEP 57637616.2.0000.5505) and all students who have answered signed an Informed Consent Form, allowing the usage and publication of their data for academic purposes. After two years of collection (2016 and 2017), it was obtained consistent data of 618 students. Students' entrances of the data range from 2010 and 2017, with a high concentration in 2015 (23.5%), 2016 (31.5%) and 2017 (14,9%). Approximately 80% of the answers were collected from students with full time classes and 95% corresponding to individuals with single marital status. Students' ages ranged from 16 to 42, but 75% of them were less than 22 years old. In order to access the main students' cultural activities, questions based on socio-economical aspects and cultural habits and interests were selected and can be visualized in Table I.

4.3 Data analysis

After collecting the data, a manual curation was made in order to remove incorrect and incomplete answers. The usage of the BRKGA algorithm considers the data represented in a complete weighted graph. Each edge of the graph represents a student and its weights (distances or dissimilarities) are

Table I. Used variables

Category	***Subcategories/Answers***
Religion	Catholicism, Evangelicalism, Afro-Brazilian Religions, Judaism, Islam, Spiritualism, Atheism, Hinduism, Buddhism, Others
Education	Private School, Public School, Private School in a scholarship program
Extra Courses	Courses for college admittance, Technical courses, Industrial training courses, Language courses, Others
Social class	A, B, C, D
Economical Situation	Not working, Working and economically dependent (partially), Working and economically independent, University financial support
Scholarship	Financial status, Scientific initiation, Others
Admittance quotas	Not used, Financial status, Skin color, Education in public school
Hobby	Art fairs, Circus, Soirée, International concerts, National concerts, Gourmet events, Concerts of classical music, Dancing performances, Reading books, Listening to music, Playing games, Watching series/movies, Going to bars, Dancing, Going to the mall, Singing, Playing a musical instrument, Writing, Sports, Crafts
Sports	Fighting sports, Soccer, Olympic Gymnastics, Volleyball, Basketball, Aquatic sports, Gym, Artistic Activities
Reading habits	Not reading, 1 to 3 books per year, 3 to 10 books per year, 10 to 20 books per year, more than 20 books per year
Frequency	Movies, Theater, Soccer stadiums, Museums, Mall, Parks, Concerts, Night clubs
Artistic Activities	Playing in a band, Playing a musical instrument, Dancing, Theater, Producing audiovisual material, Drawing/Painting, Writing, Playing games, Others

calculated using the cosine similarity function, where the identification of the relation is calculated by looking at the angle instead of magnitude. The BRKGA was implemented in C language and used with four different numbers of clusters (2, 3, 4 and 5). The used parameters were: (1) 100 individuals in the population; (2) 100 generations; (3) the size of the elite set in population was 0.2; (4) the number of mutants to be introduced in population at each generation was 0.2; (5) the probability that an allele is inherited from the elite parent was 0.6. After the identification of groups, a Principal Coordinates Analysis (PCoA) was performed in order to summarize and represent inter-object dissimilarity in a low-dimensional Euclidean space. In addition to preserve Euclidean and χ^2 distances between objects, this analysis could preserve the distances generated by the cosine similarity metric.

In order to evaluate the efficiency and detect the differences between clusters, we selected the variables with at least 50% or 25% of differences between two or more groups. After the identification of these characteristics, they were plotted in a heat map, showing the relative abundance of students in each group. The performed analyses of the study were executed in a personal computer with Intel® Core™ i7-7500U CPU @ 2.70GHz × 16 GB of memory RAM.

5. RESULTS

With the aim of analyzing the differences between the four configurations (2, 3, 4 and 5 groups) used in the BRKGA, the fig. 1 presents the results of the PCoAs performed with the different results.

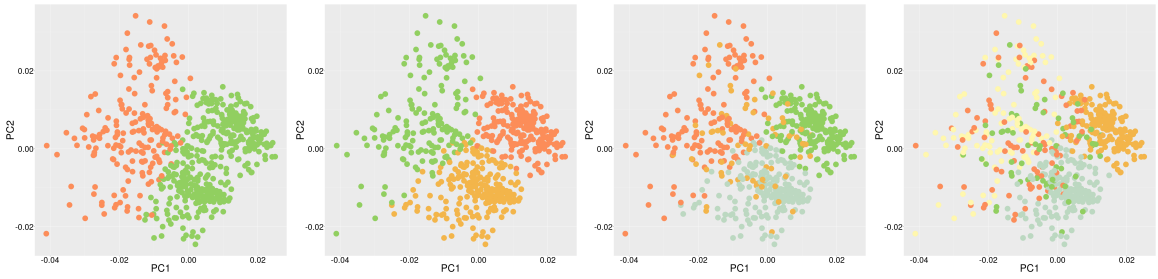

Fig. 1. Principal Coordinates Analysis (PCoA) using cosine similarity metrics; individuals are labeled according to BRKGA results using 2, 3, 4 and 5 groups.

The clear separation of groups in PCoA is seen when using 2 and 3 groups. Even though the usage of 4 and 5 clusters generate groups with high overlapping of objects, the cultural interpretation using these results showed similar results as the others. The distribution of the quantity of elements between groups in the tests was not discrepant and presented the following quantities of elements: (1) 2 groups: 424 and 194; (2) 3 groups: 168, 218 and 232; (3) 4 groups: 201, 138, 69 and 210; and (4) 5 groups: 67, 92, 182, 169 and 108.

Using two clusters and a minimum percentage of distinction between groups of 50%, there was a clear distinction of students who had as one of their favorite activities 'going or not to bars' as showed in fig. 2. The first group (1G) was composed by approximately 80% of students who had indicated this activity as a favorite one in contrast to the second group (2G) where 81% of the students answered the opposite.

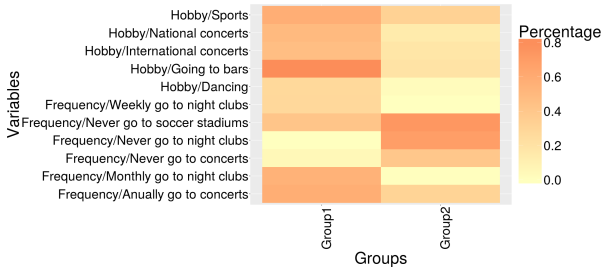

Fig. 2. Heat map of variables with at least 25% of difference between groups using two clusters.

Another interesting aspect to be noticed is that the frequency of going to night clubs corroborates the previous association. 70% of 2G answered that never go to night clubs in comparison to 1% of 1G, which had the most associated frequencies to night clubs as monthly (55%) and weekly (28%). Reducing the minimum percentage of distinction between groups to 25%, it is clear that, associated with going to bars and night clubs, there was a medium interest of dancing and going to concerts by 1G contrasted by a low interest in these activities by 2G, what was expected based on previous conclusions.

Adding one more group to the analysis, this cultural distinction remained evident. With three groups and 50% of minimum distinction percentage between groups, the contrast between students who considered night activities as hobbies continued to be clear. Two groups had approximately 77% and 80% of components with interest on going to bars (hobby) compared to 18% of the other group. With 25% of minimum distinction, these same groups showed a higher frequency on going to night clubs, theaters and concerts as it can be seen in fig. 3. One more point that ratifies this distinction is the presence of the religion variables with percent significance among groups. As expected, in groups with more protestants, night habits were not observed with as much intensity as in the groups with more catholics and atheists. It is also interesting to point that using two and three groups, people interested in sports are also interested in going to bars or nightclubs. This might be explained due to the fact of the presence of college entities associated with both sports and throwing parties. However this is not a general rule, what can be observed in fig. 2 and 3, which have a small amount of students who are not interested in night activities and are interested in sports.

With four clusters, the quantity of relevant variables with minimum 25% of difference among groups was much more superior than the other tests (25 cases). The same previous interpretations about going to bars/nightclubs might be made, which can be seen in fig. 4, where there are variables with at least 50% of difference between two groups or more. However in this case the heterogeneity between groups is evident. This quantity of clusters showed to be not appropriate to such analysis due to the fact that it would not be so meaningful when used by itself, different from the previous groups, where its significance may be easily seen without other complementary observations. The last test was

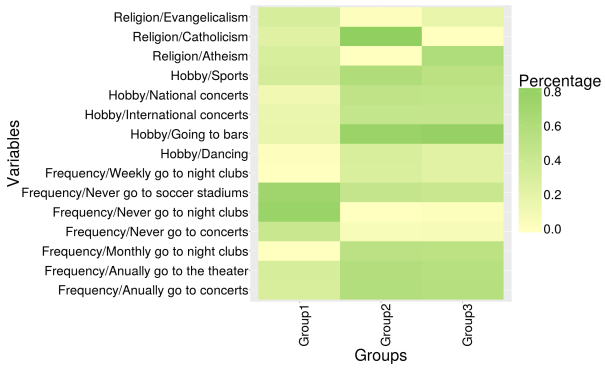

Fig. 3. Heat map of variables with at least 25% of difference between two or more groups using three clusters.

similar to the usage of four clusters. The quantity of five groups also showed to be not appropriate for this type of analysis. However, using a percent difference of 50% it could be noticed that two new variables showed significance to the distinction of groups (absence of admittance quotas and economical situation of not working). One group separated from the others due to the presence of individuals who have different sources of income and people with quotas represent two distinct groups that contrast the other three.

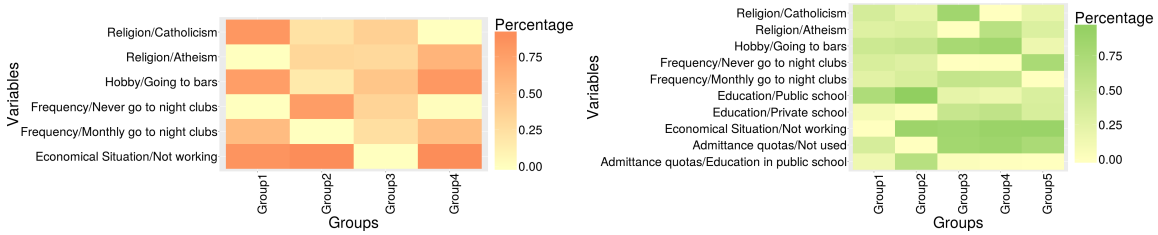

Fig. 4. Heat map of variables with at least 50% of difference between two or more groups using four and five clusters.

6. DISCUSSION

The presented results show that it is possible to identify different cultural subunits in the university. The most different aspect among groups was a high distinction of students with night activities as a common leisure practice. The usage of BRKGA with different numbers of clusters presented punctual differences, but a high segregation similarity in this characteristic. Other leisure and artistic activities did not present significant influence to the groups' formation, which is compatible to [Barth 1998], who presents the definition of culture as a discontinuous form of cultural interests. Herein the separation due to individuals' entertainment and pleasure-seeking at night time might be interpreted as the interconnected differences between students, while other interests and activities are the common culture and practices in the university.

[Chatterton and Hollands 2002] state that changes occurring within cities related to night-time economy act as one of the backdrops for understanding the cultural transformations in young people lives and in our study this concept was clearly understood when the previous interpretation is used. These results can be used by university's departments interested in promoting culture in the institute and also offer to life science researchers a different way of analyses using cultural data. Due to the fact that social interactions and not cultural interests could represent a key factor to students' separation, the promotion of social and cultural actions in the university presenting different activities might be

seen as a form of integrate these students, what is considered a key point in the whole society [Taylor 1991].

[Malbon 1998] points that the unity of identity appears to be far less significant in contemporary youth culture than has been recognized by theorists of youth culture up to now. [Chatterton and Hollands 2002] on the other hand emphasize the idea of how factors like educational background, parental income, and ethnicity are related to nightlife consumption practices, indicating aspects of how it can exclude groups. Having the idea that organizational culture can influence the construction of identity and the fact that the insertion of people into cultural groups provides modes of human behavior [Wang 2007], the study of these groups' cultural distinctions brings a new way of understanding the significance of contemporary segregation and how it is linked to cultural habits and ways of living in the context of the university.

7. CONCLUSION

This paper presented an analysis of unsupervised clustering using BRKGA and cultural data. We achieved positive results by separating groups of students with clear distinguishable cultural characteristics, allowing the identification of particular modes of social interaction. Since the contexts of social interactions are key factors in the opportunities to refashion ourselves and identify with others [Malbon 1998], the identification of different groups and the creation of university's actions to promote students' integration and more access to different forms of culture is extremely important to students' formation and expansion of their cultural identity. Understanding how the individuals are organized is the first step to comprehend the segregation in the institute and what could be done to reduce it.

Acknowledgment

The authors would like to acknowledge the Extension and Culture Chamber (CaEC) of Federal University of São Paulo and the Support Student's Nucleus (NAE) of the institute.

REFERENCES

Babaki, B., Guns, T., and Nijssen, S. Constrained clustering using column generation. In *International Conference on AI and OR Techniques in Constriant Programming for Combinatorial Optimization Problems*. Springer, pp. 438–454, 2014.

Barth, F. *Ethnic groups and boundaries: The social organization of culture difference*. Waveland Press, 1998.

Bean, J. C. Genetic algorithms and random keys for sequencing and optimization. *ORSA journal on computing* 6 (2): 154–160, 1994.

Canclini, N. G. *Consumers and citizens: Globalization and multicultural conflicts*. Vol. 6. U of Minnesota Press, 2001.

Chatterton, P. and Hollands, R. Theorising urban playscapes: producing, regulating and consuming youthful nightlife city spaces. *Urban studies* 39 (1): 95–116, 2002.

Gonçalves, J. F. and Resende, M. G. Biased random-key genetic algorithms for combinatorial optimization. *Journal of Heuristics* 17 (5): 487–525, 2011.

Kroeber, A. L. The concept of culture in science. *The Journal of General Education* 3 (3): 182–196, 1949.

Malbon, B. Clubbing: consumption, identity and the spatial practices of every-night life. *Cool places: Geographies of youth cultures*, 1998.

Merriam-Webster Dictionary. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 2002.

Morgado, A. C. As múltiplas concepções da cultura. *Múltiplos Olhares em Ciência da Informaç* 4 (1), 2016.

Resende, F. G. and de Paula, A. V. Influência da cultura organizacional na (re) construção da identidade dos trabalhadores: um estudo de caso em uma empresa de tratamento de resíduos no sul de minas gerais. *Psicologia: teoria e prática* 13 (3), 2011.

Spears, W. M. and Jong, K. D. D. On the virtues of parameterized uniform crossover. Tech. rep., NAVAL RESEARCH LAB WASHINGTON DC, 1995.

Taylor. The multicultural organization. *The executive*, 1991.

Wang, Y. Globalization enhances cultural identity. *Intercultural Communication Studies* 16 (1): 83, 2007.