

Expressive and semantic description of human capital diversity: a machine learning approach

Roberto Douglas Guimarães de Aquino

Federal University of São Paulo (UNIFESP)

Avenida Cesare Mansueto Giulio Lattes, 1201, São José dos Campos - SP, Brazil

aquinordga@gmail.com

Vitor Venceslau Curtis

Aeronautics Institute of Technology (ITA)

Praca Marechal Eduardo Gomes, 50, São José dos Campos - SP, Brazil

curtis@ita.br

Filipe Alves Neto Verri

Aeronautics Institute of Technology (ITA)

Praca Marechal Eduardo Gomes, 50, São José dos Campos - SP, Brazil

verri@ita.br

ABSTRACT

Human capital (HC) and its diversity are valuable assets of organizations, contributing to the longevity of companies and the success of startups. Recently, the industry has employed machine learning (ML) techniques in human resource management (HRM) to automate people selection, but considering mainly their individual technical skills. Conversely, to obtain better team performance and avoid social bias, complementary capacities and diversity of HC are desired. Analyzing the HC diversity imposes many challenges, including its semantic description. Studies in HRM propose to solve the identification of HC diversity using clustering methods, but analyzing their result may be impractical because of the large number of features of current datasets. This work presents an automatic pipeline that chooses the best clustering method, and semantically describes clusters of employees learned with the minimum of external input, leaving the decision-makers only with the task of providing the candidates' dataset and the number of desired partitions. We introduce a semantic descriptor operator $\star$ to describe the characteristics of clusters, filtering only those statistically relevant, allowing easy interpretation and potential information to assist decision-making. Finally, we validate our pipeline with a real dataset to present its semantic potential.

KEYWORDS. Descriptive learning. Diversity. Human Resources Management.

1. Introduction

According to Bates [1990]; Clough et al. [2019], human capital is one of the most valuable assets for organizations specially with regard to their survival, longevity, quality, efficiency and to attract investments. The concept, originally described on economics by Becker [1964] (Nobel Prize) means an intangible asset useful in the production process but owned by the employee, e.g., their knowledge, skills, abilities, health, and values.

Since Becker [1964], organizations have been invested heavily in human capital (e.g. education, training, health care, etc.) as a key strategy to increase company performance, but it is only effective when accompanied by efficient human resource management (HRM). The research field of strategic HRM has been studied how the management of human resources within organizations influences the organizational and individual outcomes to increase the firm performance, as described by Boon et al. [2018].

A major discovery of this research area is the diversity as an important driver for the success of economic activities, i.e., a factor of the human capital derived from all other factors of an organization. Recent studies, as Meili and Shearmur [2019]; Chesbrough and Brown [2006]; Huizingh [2011], have evinced the importance of diversity, specially in small towns and rural areas. Companies from such areas lack from diversity but they have been effectively succeeded in innovation and competition by increasing diversity through collaborations with external partners.

This contrasts to the passive investments in the human capital within the organization to improve performance in the sense that you can benefit from the diversity of collaborations but without direct investments. The trade-off is the non-exclusivity of the shared human capital.

In this work, due to the complexity of the problem, we focus on the first step of this strategic HRM process. We propose a machine learning (ML) framework to identify and profile the diversity a company holds. The framework is able to identify the statistically distinct groups of human capital and report a semantic description of these strategical groups in a meaningful and semantic format to support human resource experts and managers in decision making.

Currently, the HRM industry has employed ML to address strategic HRM concerns as in Strohmeier and Piazza [2013], including candidate selection as Aioli et al. [2009] suggests and diversity as described by Akgün [2020]; Sipior [2020]. Some other examples are identification of competencies in Zhu et al. [2005], assignment of employee to tasks in Bohloul et al. [2017], prediction in Thissen-Roe [2005] and evaluation in Zhao [2008a,b] of employee performance.

Our framework is general to be applied in other typical clustering applications to automatically identify and describe the semantic meaning of clusters. The rest of this paper is organized as follows. Section 2 briefly discuss about how ML can be applied in strategic HRM. Section 3 introduces the basic concepts and methods used throughout the article. Section 4 introduces the proposed model. Section 5 describes the experiments and results using a real dataset of HRM, and Section 6 has the conclusions and future works.

2. Strategic human resource management

Although the final goal is to achieve better decisions for a company's human resource, it starts with the understanding its human capital. For example, HRM experts performing task assignments and people selection can benefit HRM models by simply restricting the choices among the options pre-validated by HRM experts. Thus, the first challenge is to identify which features of the human capital to analyze, e.g., what abilities and skills to measure. In this sense, strategic human capital (SHC) is a recent research field that studies how human capital may serve as a valuable resource and how these skills and knowledge can be captured and transferred, as described by Boon et al. [2018].

With the desired features validated by the HRM and managers, another typical problem for HRM is performance prediction. The outcome of a group of people performing any kind of task depends not only on the abilities and skills of individuals, but also on the complementary differences among the individuals, resulting in a complex result to predict. Thus, choosing the right features to describe and understand employee performance plays a key role in strategic HRM. For this problem, we suggest ML as a tool to help HRM better understand the human capital of an organization.

Finally, it is also important to mention the operational problems that hinder an efficient assignment of human resources to accomplish tasks and meet goals: it is a time-consuming task, typically involves manual surveys and interviews, expectations of experts must be aligned with the HRM, the quality of the assignment is proportional to the amount of people analyzed, and others. Thus, it becomes impractical to perform these assignments efficiently without automation.

3. Concepts and Methods

In this section, we briefly describe the two clustering approaches explored in our model and the concept of *itemset* used to describe the revealed distinct groups. In this work, we consider the terms *cluster* and *group* equivalent.

3.1. K-means Clustering

Let $X = \{x_i : i = 1, \dots, N\}$ be the set of N D -dimensional points (or people's features) to be grouped into a set of K disjoint clusters, $C = \{c_k : k = 1, \dots, K\}$. K-means algorithm finds a partition such that the distance between the centroid of each cluster and the points in the cluster is minimized. Let

$$\mu_k = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$$

be the centroid of the cluster c_k , K-means searches for a partition that minimizes

$$J = \sum_{k=1}^K \sum_{x_i \in c_k} d(x_i - \mu_k),$$

where d is an arbitrary metric, using a greedy iterative heuristic, as described by Jain [2010].

3.2. Hierarchical Clustering

According to Duda et al. [2000], an agglomerative hierarchical clustering algorithm generates a sequence of partitions of N points in which the first partition presents N groups of one point each and the last one presents one group with all points. Thus, the data is grouped in such a way that, if two points are grouped at some level, at the higher levels they remain part of the same group, building a hierarchy of clusters.

To generate such a sequence, starting from the first partition, clusters are merged using a strategy, called linkage criteria, that depends on an arbitrary metric to measure the distance between data points.

3.3. Frequent Itemsets

Let T be a set of transactions, where each transaction contains a set of items. Items can be products, components, services, skills, etc. And a transaction can be represented by a set of products purchased in a supermarket, a set of pages visited in a section, a set of skills identified in a profile, etc. A set of items is also called an *itemset*. An itemset with k items is called a k -itemset. The *support* of an itemset is the number of transactions in T that it occurs as a subset. A k length subset of an itemset is called a k subset. An itemset is maximal if it is not a subset of any other itemset. An itemset is *frequent* if its support is more than a user-specified *minimum support* value, as defined by Zaki [2000].

4. Model Description

According to our proposal, the decision maker inputs a dataset and the number of partitions K one desires to group the data. The dataset consists of a simple table with employees by rows and their features by columns. At the end, a report is generated with the results containing each of the groups described through intrinsic frequency patterns. All steps in the pipeline are executed automatically. The fig. 1 presents an overview of the model and each of its steps, detailed in the following sections.

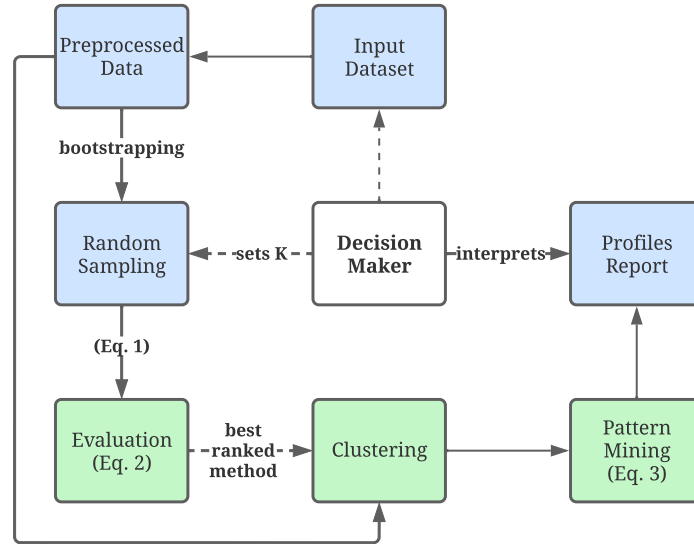


Figure 1: Flowchart describing each of the stages of the pipeline proposed below. Blue boxes represent data and green boxes represent processes. Solid arrows represent process flow and dashed arrows represent configuration settings. The decision maker loads the input data and informs the number of groups through the knowledge of the cause. The algorithm automatically processes the data, evaluates the best algorithm, performs the clustering, searches for patterns in the results and generates a report with the identified profiles.

4.1. Dataset Preprocessing

Let $\tilde{X} = \{\tilde{x}_i\}_{i=1,\dots,N}$ be the input dataset where $\tilde{x}_i \in \mathbb{R}^D$ represents the capabilities or characteristics of i -th person. Each component $\tilde{x}_{ij} \in \mathbb{R}$ represents how well the i -th person masters the j -th capability (alternatively, how well it performs the j -th activity, or how much of j -th characteristic is present.)

For the preprocessing step, it is important to select the most significant features for the profile definition, remove missing values and duplicate observations. The preprocessed dataset, $X = \{x_i\}_{i=1,\dots,N}$, is used in the next steps of the pipeline.

4.2. Clustering of Bootstrapped Subsets

The next step of the pipeline is to measure the performance of different clustering algorithms in the dataset X . We want to select the algorithm that not only achieves the best expected performance but also has lower variance.

Given a dataset O , a clustering algorithm a , and the partition evaluation function s , we define $a(O)$ the K partitions of O obtained by a and $s(a(O))$ the performance of a in O in terms of s .

Let s be an arbitrary evaluation function, $X_1, \dots, X_P$ be sets obtained by random sampling M elements of X with replacement, and $a_1, \dots, a_Q$ be different clustering algorithms, we calculate the matrix S whose elements are $s_{pq} = s(a_q(X_p))$.

We require that not only the algorithms but also the evaluation function use the same metric d to calculate the distance between data points. In the following step in the framework, we use matrix S to select one of the clustering algorithms.

4.3. Clustering Evaluation and Algorithm Selection

In data classification, it is a well-accepted practice to compare the performance of classifiers using the Friedman test and the Nemenyi post-hoc test Demšar [2006].

In clustering, although present in literature such as Hatamlou [2013], their application is not common. Since we can assume the sets X_p are independent between themselves for all p as in Davison and Hinkley [1997] the set A of non-dominated algorithms is

$$A = \left\{ a_q : \nexists q' \neq q \text{ such that } \frac{1}{P} \sum_p r_{pq'} < \frac{1}{P} \sum_p r_{pq} - \text{CD} \right\}, \quad (1)$$

where r_{pq} is the rank of s_{pq} in the population $\{s_{pq'} : \forall q'\}$ and CD is the critical difference of the Nemenyi post-hoc test with significance α , as described in Demšar [2006] (We assume that the Friedman test rejects the null-hypothesis that the performance measurements in matrix S are equivalent¹).

We select the algorithm $a_{\text{best}} \in A$ that yields the lowest performance variation, that is,

$$a_{\text{best}} = \underset{a_q \in A}{\text{argmin}} \text{var} \left(\{s_{pq}\}_{p=1, \dots, P} \right), \quad (2)$$

where $\text{var}(\cdot)$ is the sample variance. We believe that if the methods are statistically equivalent, the most stable one is preferred.

For the next step of the pipeline, we consider the partition $\{c_1, \dots, c_K\} = a_{\text{best}}(X)$, where each c_k is named a cluster.

4.4. Semantic Description of the Groups

In this section, we introduce a non-associative operator $\star$ to semantically describe the groups, and use the *dominating frequent patterns* of the clusters to describe, respectively, unique or frequent elements to provide a semantic summarization of the important features of a cluster.

To semantically describe the groups, it is trivial to transform a set of data points in a set of transactions. Let T be the set of transactions derived from the set X after discretizing its data points, we define $F_1(O)$ as the collection of all frequent itemsets with length 1 in T given a minimum support $\text{sup}_{\min}$.

We define the dominating frequent patterns $E(c_k)$ of the cluster c_k as

$$E(c_k) = \left(\left((F_1(c_k) \star F_1(c_1)) \star F_1(c_2) \right) \star \dots \right) \star F_1(c_K), \quad (3)$$

where $F \star F'$ yields a collection of itemsets whose elements are either

- present in F but not in F' ; or
- present in both F and F' but its support in F is not less than its support in F' .

¹If the Friedman test does not reject the null-hypothesis that the performance of algorithms are the same, we set A comprises all algorithms.

4.5. Remarks on the Parameters

Our proposed framework has a few parameters that should not depend on the user input. Table 1 summarizes them and indicates sane default values.

Parameter	Suggested value
Number of bootstrapped sets (P)	15
Size of the bootstrapped sets (M)	$\left\lceil \frac{-\log N}{P \log(1 - 1/N)} \right\rceil$
Statistical test significance (α)	0.05
Minimum support ($\text{sup}_{\min}$)	0.5
Metric (d)	Euclidean distance
Partition evaluation measures (s)	Average silhouette score Rousseeuw [1987]

Table 1: Parameters of the proposed framework and their suggested values.

Our recommendation for α is a conservative value used in many research works. The suggested size M guarantees that the expected number of data items that do not belong to any of the bootstrapped sets is less than 1. Also, setting $\text{sup}_{\min} = 0.5$ filters itemsets that happen in less than half of the data items in the cluster.

The chosen metric and partition evaluation measure are also common in clustering research. We further discuss alternatives in section 6. Moreover, the choice of the algorithms depends on the available computational power since it is the most time-consuming part of the pipeline. In this paper, we use K-mean and hierarchical clustering.

5. Experimental Results

The data used in the demonstration are in the public domain and were generated for teamwork management applications by Yu et al. [2017]. This database contains detailed data that reflects situations, strategies and decision results, in addition to personal data of 1,144 participants from different origins.

5.1. Dataset Preprocessing

In the pre-processing phase, we chose the features most related to the personality of the participants, in order to test the application of the method and identify different profiles. In this way, we select the most relevant features from the database: *Extraverted*, *Critical*, *Dependable*, *Anxious*, *Open to new experience*, *Reserved*, *Sympathetic*, *Disorganized*, *Calm* and *Conventional*. The values in each of them are of the ordinal, integer quantitative type, belong to the range between 1 and 5 and are related to the participants' responses. In this way, each answer can be related to one of the categories: 1 - *Disagree strongly*, 2 - *Disagree*, 3 - *Neither agree nor disagree*, 4 - *Agree*, and 5 - *Agree strongly*. Finally, after selecting the features, the observations with missing values are removed and we obtain a database with 161 examples. The fig. 2 shows the distributions of each of the features by histograms.

5.2. Clustering of Bootstrapped Subsets

The purpose of this step is to measure the performance of different clustering algorithms in the X data set. For this, we consider all 161 examples (N) from the database and a number of bootstrap sets (P) equal to 15 with sizes equal to

$$M = \frac{-\log N}{P \log(1 - 1/N)}, \quad (4)$$

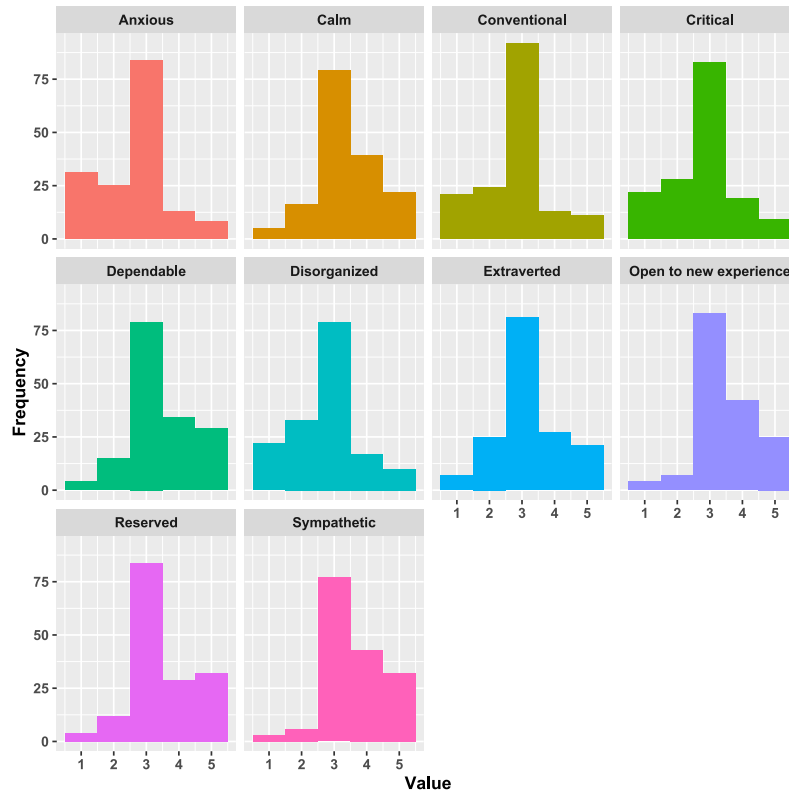


Figure 2: Representation of the distributions of each feature through histograms.

Subset	kmeans	ward	average	single	complete
1	0.277	0.336	0.285	0.248	0.176
2	0.440	0.413	0.314	0.274	0.305
3	0.340	0.380	0.231	0.185	0.290
4	0.376	0.409	0.420	0.336	0.372
5	0.446	0.425	0.445	0.421	0.432
6	0.336	0.421	0.318	0.276	0.376
7	0.433	0.476	0.372	0.376	0.369
8	0.275	0.315	0.404	0.341	0.305
9	0.393	0.449	0.465	0.310	0.357
10	0.315	0.358	0.253	0.253	0.279
11	0.398	0.387	0.380	0.338	0.332
12	0.429	0.410	0.315	0.152	0.395
13	0.340	0.359	0.366	0.282	0.327
14	0.391	0.408	0.362	0.343	0.385
15	0.380	0.406	0.405	0.392	0.370

Table 2: Matrix S containing the silhouette scores for P subsets calculated for each proposed method.

as proposed in table 1, to calculate the matrix S . The matrix S is presented table 2 in table 2.

The number of partitions K defined is 6 and was chosen empirically. As defined in the

description of the pipeline, presented by fig. 1, the decision maker is expected to inform the number of groups, based on domain knowledge.

For the next step, the S matrix is used to select one of the clustering algorithms.

5.3. Clustering Evaluation and Algorithm Selection

Thus, we can compare the performance of the algorithms (A) using the Friedman test and the Nemenyi post-hoc test. The table 3 shows the ranks (r_{pq}) calculated for each of the P subsets using each proposed methods.

Subset	kmeans	ward	average	single	complete
1	3	1	2	4	5
2	1	2	3	5	4
3	2	1	4	5	3
4	3	2	1	5	4
5	1	4	2	5	3
6	3	1	4	5	2
7	2	1	4	3	5
8	5	3	1	2	4
9	3	2	1	5	4
10	2	1	4	4	3
11	1	2	3	4	5
12	1	2	4	5	3
13	3	2	1	5	4
14	2	1	4	5	3
15	4	1	2	3	5

Table 3: Ranks (r_{pq}) calculated for each of the P subsets using each proposed methods.

As result, we can infer by the critical difference diagram (CD Diagram), fig. 3, that there are no significant performance differences between the *kmeans* and *ward* methods (Hierarchical Clustering).

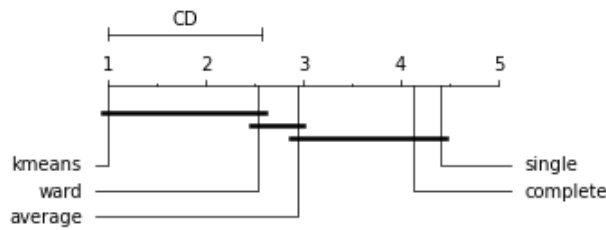


Figure 3: Critical difference diagram (CD Diagram) relating the ranks of the methods.

However, the *ward* method has less variance, as can be seen in the table 4. Thus, using the selection criteria defined in section 4, *ward* can be configured as the best method (a_{best}).

5.4. Semantic Description of the Groups

After evaluating and defining the best method, the algorithm is applied again in the database and the frequent patterns of each group are obtained. As described in table 1, we use a minimum support of 0.5. Thus, frequent items that are observed in less than 50% of the group

Method	Standard deviation
kmeans	0.0560
ward	0.0407
average	0.0661
single	0.0715
complete	0.0597

Table 4: Variance of the S matrix with the silhouette scores for each method.

members are ignored. Given that there are 5 categories, the probability of a feature having one of them is 20%, considering a uniform probability distribution as a reference for randomness in filling in the features of the members, we believe that this minimum support value is adequate as a threshold.

Finally, a report is generated with the groups' frequent dominant patterns, obtained after applying eq. (3) to the identified frequent itemsets. table 5 shows the count of frequent dominant patterns in each of the groups.

Cluster	1	2	3	4	5	6
Number of dominant frequent patterns	10	7	3	3	2	5
Proportion of members [%]	56.52	11.18	5.59	5.59	12.42	8.69

Table 5: Number of frequent dominant patterns and proportion of members within each of the groups.

In the next section, we present and discuss the profiles identified in each of the groups, based on the semantic descriptions obtained in the final report.

5.5. Discussion of Results

In this section, we briefly discuss the final report containing the semantic description of each of the groups, presented by fig. 4. Before, we emphasize the importance of using pattern-based analysis, especially when dealing with categorical data, since statistical analyzes aimed at identifying a semantic description of groups can be imprecise or inadequate.

Given that there is also diversity within groups, which is lower among members of the same group when compared to others, the frequent dominant patterns are identified proportionally in each group. This proportionality is defined by the support value, which can vary from each other. We also emphasize that not all personality characteristics can appear in the semantic description of the groups, since the method selects the most relevant features.

Observing the results referring to group 01, presented by fig. 4a, we verified that the identified profile has a distribution equal to 3 (Neither agree nor disagree) in all 10 features, indicating that most of the group members have doubts about the aspects that define your personality. As for the proportion of members that are described, we found that between 71% and 78% have this question regarding the following aspects: Calm, Dependable, Critical, Disorganized, Extraverted, Anxious, Reserved and Sympathetic, in ascending order of percentage. Thus, for the Open to new experience and Conventional aspects, proportions of described members of 80% and 83% are associated, respectively.

Regarding group 02, presented by fig. 4b, we found that between 50% and 55% of the members strongly believe (5 - Agree strongly) that they are calm, open to new experiences and extroverted. As for Conventional and Disorganized aspects, between 55% and 66% of the members

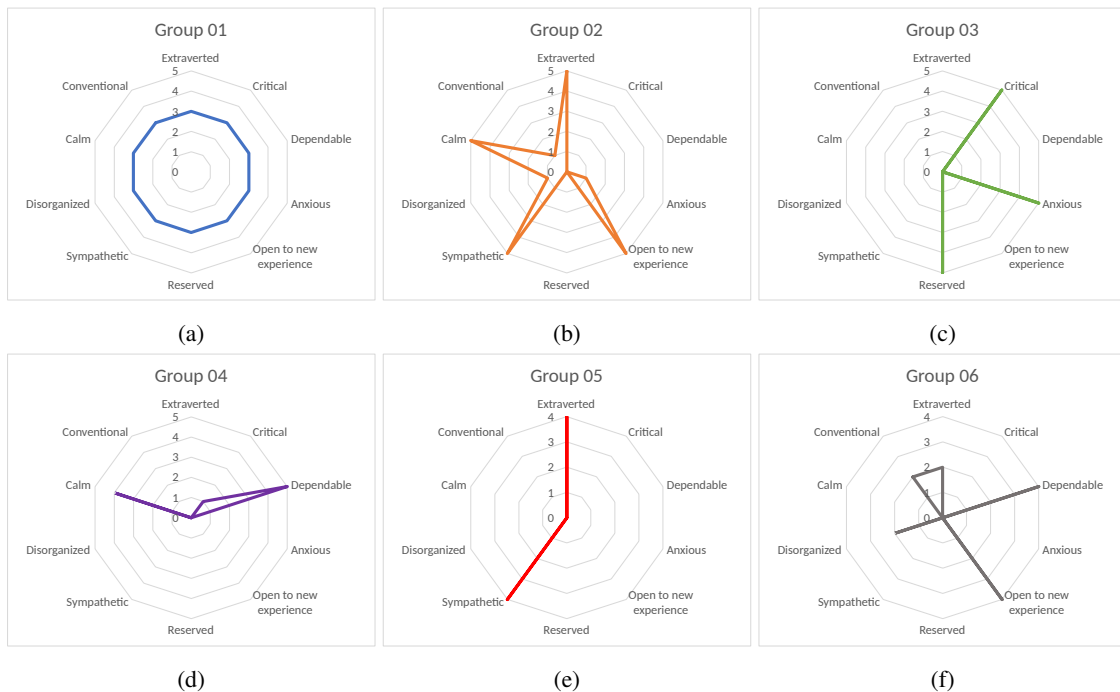


Figure 4: Representation of profiles based on the semantic description of each of the groups.

were identified. Finally, 72% strongly agree that they are sympathetic and 77% strongly disagree that they are anxious.

On the other hand, in group 03, presented by fig. 4c, 55% of the members feel anxious, 66% consider themselves critical and 88% believe they have a more reserved profile. Meanwhile, the members of group 04, presented by fig. 4d, are described: 66% strongly believe they are calm and reliable; and 55% do not recognize themselves as critical.

Group 05, presented by fig. 4e, contains members in which 70% of them believe they are extroverted and 60% that they are sympathetic. Finally, in group 06, represented by fig. 4f, 50% believe that they are reliable and open to new experiences. However, about 57% disagree that they are conventional, extroverted and disorganized.

As we can see from the figures and descriptions, the groups obtained by the clustering algorithm identified different profiles, each with a unique set of characteristics. Considering the diversity between and within groups, we found that the semantic description cannot define in an absolute way the profile of the members within the group, however, we believe that it can be used as an important tool to aid decision making.

6. Conclusion

This work introduces an HRM framework based on ML techniques to automatically identify groups (or clusters) of similar people with the minimum of external input, requiring no statistical or ML skills. Our framework can statistically check among a list of clustering techniques the one with the best clusters and produce a small report about the semantic meaning of each cluster.

The empirical results presented in section 5 with a real dataset of HRM confirm the potential of conciseness and easy interpretability of the final report of our framework, obtained by the proposed semantic descriptor operator $\star$ which filters the features of an identified cluster to only statistical and meaningful information to understand each cluster in contrast to the others. We expect

that this approach can speed up the tedious and manual work of interpreting the identified cluster made by domain experts and, thus, also speed up decision-making. To the best of our knowledge, our proposal is the first attempt to address the posterior analysis problem of clustering methods. Thus, we also expect to motivate the research of new innovative ways to help the automation of this tedious work.

Although our framework is general for other applications, in this work we focus on HRM due to the potential impacts it may have for the society and economies. We believe that models in strategic HRM that takes into account human capital diversity can achieve better team performance and reduce the social bias problem, e.g., by restricting the feasible choices of the search space of black box models with social bias problem to choices with diversity.

For future work, we also identified some points that could be better developed, such as some improvements to the $\star$ operator, use of other distance metrics, different weights for attributes, use of other clustering algorithms, applications in other databases and heuristics to identify the best support value to find more relevant descriptions.

7. Acknowledgments

This study was financed in part by the Brazilian Federal Foundation for Support and Evaluation of Graduate Education (CAPES) - Finance Code 001.

References

- Aioli, F., De Filippo, M., and Sperduti, A. (2009). Application of the preference learning model to a human resources selection task. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, p. 203–210. IEEE.
- Akgün, A. E. (2020). Team wisdom in software development projects and its impact on project performance. *International Journal of Information Management*, 50:228–243.
- Bates, T. (1990). Entrepreneur human capital inputs and small business longevity. *The review of Economics and Statistics*, p. 551–559.
- Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press, Chicago.
- Bohlouli, M., Mittas, N., Kakarontzas, G., Theodosiou, T., Angelis, L., and Fathi, M. (2017). Competence assessment as an expert system for human resource management: A mathematical approach. *Expert Systems with Applications*, 70:83–102. ISSN 0957-4174.
- Boon, C., Eckardt, R., Lepak, D. P., and Boselie, P. (2018). Integrating strategic human capital and strategic human resource management. *The International Journal of Human Resource Management*, 29(1):34–67.
- Chesbrough, H. W. and Brown, J. S. (2006). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business School Press, Boston.
- Clough, D. R., Fang, T. P., Vissa, B., and Wu, A. (2019). Turning lead into gold: How do entrepreneurs mobilize resources to exploit opportunities? *Academy of Management Annals*, 13(1): 240–271.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge, United Kingdom.

- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*. Wiley, New York.
- Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information sciences*, 222:175–184.
- Huizingh, E. K. R. E. (2011). Open innovation: State of the art and future perspectives. *Technovation*, 31(1):2–9. ISSN 0166-4972. Open Innovation - ISPIM Selected Papers.
- Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8): 651–666.
- Meili, R. and Shearmur, R. (2019). Diverse diversities—open innovation in small towns and rural areas. *Growth and Change*, 50(2):492–514.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Sipior, J. C. (2020). Considerations for development and use of ai in response to covid-19. *International Journal of Information Management*, 55:102170.
- Strohmeier, S. and Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, 40(7):2410–2420. ISSN 0957-4174.
- Thissen-Roe, A. (2005). *Adaptive selection of personality items to inform a neural network predicting job performance*. PhD thesis, University of Washington.
- Yu, H., Shen, Z., Miao, C., Leung, C., Chen, Y., Fauvel, S., Lin, J., Cui, L., Pan, Z., and Yang, Q. (2017). A dataset of human decision-making in teamwork management. *Scientific Data*, 4(1): 1–12.
- Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(23):372–390.
- Zhao, X. (2008a). An empirical study of data mining in performance evaluation of hrm. In *2008 International Symposium on Intelligent Information Technology Application Workshops*, p. 82–85. IEEE.
- Zhao, X. (2008b). A study of performance evaluation of hrm: based on data mining. In *2008 International Seminar on Future Information Technology and Management Engineering*, p. 45–48. IEEE.
- Zhu, J., Goncalves, A. L., Uren, V. S., Motta, E., and Pacheco, R. (2005). Mining web data for competency management. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, p. 94–100. IEEE.