BOHR International Journal of Internet of things, Artificial Intelligence and Machine Learning
2023, Vol. 2, No. 1, pp. 11–18
DOI: 10.54646/bijiam.2023.12
www.bohrpub.com

METHODS

Discrete clusters formulation through the exploitation of optimized k-modes algorithm for hypotheses validation in social work research: the case of greek social workers working with refugees

Alexis Lazanas[1]*, Ilias Siachos[1], Dimitra-Dora Teloni[2], Sofia Dedotsi[2] and Aristeidis G. Telonis[3]

[1]Department of Mechanical Engineering and Aeronautics, University of Patras, Rion-Patras, Greece
[2]Department of Social Work, University of West Attica, Athens, Greece
[3]Department of Human Genetics, Miller School of Medicine, University of Miami, Coral Gables, FL, United States

***Correspondence:**
Alexis Lazanas,
alexlas@upatras.gr

Received: 21 January 2023; **Accepted:** 07 February 2023; **Published:** 18 February 2023

This article focuses on the results of self-funded quantitative research conducted by social workers working in the "refugee" crisis and social services in Greece (1). The research, among other findings, argues that front-line professionals possess specific characteristics regarding their working profile. Statistical methods in the research performed significance tests to validate the initial hypotheses concerning the correlation between dataset variables. On the contrary of this concept, in this work, we present an alternative approach for validating initial hypotheses through the exploitation of clustering algorithms. Toward that goal, we evaluated several frequently used clustering algorithms regarding their efficiency in feature selection processes, and we finally propose a modified k-Modes algorithm for efficient feature subset selection.

Keywords: clustering, k-Modes algorithm, social work, machine learning, refugee crisis

1. Introduction

Since social workers have been on the "front-lines" (2) of the so-called refugee "crisis," facing a series of difficulties in effectively helping their users. While Greece is one of the "entrance" countries in Europe, there has been no current research in social work practice with refugees. The study was a self-funded, quantitative research project with main research questions concerning, among others, the exploration of front-line professionals' profiles. Analytically, information about the research concerning: (i) aims and hypotheses, (ii) sampling strategies and research ethics, (iii) statistical methods and analysis, and (iv) the research's results can be retrieved from (1, 3).

In this article, we present a model-based approach in order to provide an alternative validation to the globally acknowledged hypotheses testing method. More specifically, several hypothesis-testing cases are commonly performed through the application of either the chi-square (4) or hypergeometric test (5) in order to determine statistical significance. This procedure is typical for discovering "correlation" between independent variables in datasets with categorical values. Augmenting this typical approach, we propose a new formally structured model by using classification techniques in order to formulate data clusters capable of validating initially composed statistical hypotheses.

To accomplish the above goal, we adopted a rather "typical" technique regarding the feature selection process from our model's data. A candidate features sub-set is created each time in order to categorize data points into intuitively similar but not predefined user groups. Then, the clustering phase is implemented through the evaluation of the following clustering algorithms: (1) a modified k-Modes

algorithm (6), (2) agglomerative clustering (7) and (3) a normal k-Modes algorithm (8) in order to adopt the most feature-selection efficient one.

The remainder of the article is organized as follows: Section "2 Related work"discusses related work issues considered in the context of literature; Section "3 Social workers' profile: Validating our initial hypothesis" refers to the statistical hypothesis testing prior to our approach, which is described in Section "4 Our approach", by demonstrating the applicability of the proposed approach through analytical steps; and finally, concluding remarks and future work directions are outlined in Section "5 Conclusion."

2. Related work

Clustering and hypothesis testing are two powerful techniques used in data analysis (9). Clustering is the process of grouping similar objects together based on their characteristics, while hypothesis testing is a statistical technique used to test the validity of a hypothesis. In recent years, researchers have combined these two techniques to gain deeper insights into their data. In this review, we will discuss the combination of clustering and hypothesis testing and its applications in different fields (10–13).

One of the primary applications of clustering and hypothesis testing is in biology, where interesting reviews can be found in Jacques and Preda (14); Wang et al. (15, 16, 17), as well as in medical imaging data (18, 19). Researchers in this field use these techniques to identify groups of genes that are related to a particular disease (20) and chemometrics (21, 22). Clustering is used to group genes that have similar characteristics (23), such as gene expression levels (24), while hypothesis testing is used to test whether the genes in each cluster are associated with the disease (25). By combining these techniques (26), researchers can identify clusters of genes that are statistically significant and associated with the disease (27).

Another area where the combination of clustering and hypothesis testing is used is in finance. Researchers in this field use clustering to group stocks that have similar characteristics (28, 29), such as market capitalization, price-to-earnings ratio, and dividend yield. Hypothesis testing is used to test whether the stocks in each cluster have significantly different returns. This can help investors identify stocks that are undervalued or overvalued, and make more informed investment decisions.

In the field of marketing, clustering methods (30, 31) and hypothesis testing are used to segment customers into different groups based on their characteristics (32) and test whether these groups have different purchasing behaviors. For example, a company may use clustering to group customers based on their age, income, and purchasing history (10). Hypothesis testing can then be used to test whether these groups have different purchasing behaviors, such as buying more frequently or spending more money. This can help companies develop more targeted marketing strategies and improve their overall sales.

In the field of image processing, clustering and hypothesis testing are used to segment images into different regions based on their characteristics and test whether these regions have different properties. For example, a researcher may use clustering to segment an image into regions based on color or texture (33, 34). Hypothesis testing can then be used to test whether these regions have different properties, such as brightness or contrast. This can help researchers better understand the properties of the image and develop more advanced image processing algorithms.

One of the primary advantages of the combination of clustering and hypothesis testing is that it allows researchers to identify statistically significant groups of data that may not be apparent using either technique alone. Clustering can help identify groups of data that are similar (35), while hypothesis testing can help determine whether these groups are statistically significant. By combining these techniques, researchers can gain a deeper understanding of their data and develop more accurate models (36).

In conclusion, the combination of clustering (37) and hypothesis testing is a powerful technique that has numerous applications in different fields (14, 38, 39). By using clustering to group similar data and hypothesis testing to test whether these groups are statistically significant, researchers can gain deeper insights into their data and develop more accurate models (40, 41). This technique has been used successfully in biology, finance, marketing, and image processing, among other fields (42, 43), and is likely to continue to be an important tool in data analysis in the future.

3. Social workers' profile: validating our initial hypothesis

A self-completed, anonymous electronic questionnaire was available online from June until September 2018, containing 52 questions (1, 3) designed by the researchers. Out of a total of 158 responses, 21 were incomplete and were thus excluded. We analyzed the 137 complete responses, and statistical significance was evaluated in R version 3.4.3, and the results' graphs were created using Excel 365 Pro Plus. We employed the hypergeometric or chi-squared test, and the statistical threshold was set at a P-value of 0.05. In this section, the findings of our research concerning the profile of social workers in the field of social services in Greece are presented in order to gradually construct our main hypothesis. To briefly describe the above concept, we note that the discussion in Authors' own (1) "provided important insights on the challenges and difficulties that social work professionals face in helping effectively their users." One of the main findings of this research was that social workers

working in the refugee "crisis" are young graduates with limited work experience. We define our main hypothesis (H_0) as follows:

H_0 = "social workers are young graduates with limited work experience."

As shown in **Figure 1**, regarding the social workers' profile, the vast majority (80%) are women and 18.3% are men. In total, 52% are between 22 and 30 years old, while 39% are between 31 and 39 years old. It is straightforward that there is a plethora of under-middle-aged professionals working on the "front-lines" with refugees, and consequently, the first part of H0 is considered undoubtedly validated.

Apart from the fact that the profession of social work is considered "female dominated", as presented in **Figure 1**, another critical point concerning social workers' overall work experience arises from the research. More specifically, as presented in **Figure 2**, there is an impressive "wide area" (almost 90%) of the professionals working on the "front line" with limited working experience varying in the interval between 0 and 4 years, within a statistically significant threshold (P-value < 0.05; hypergeometric test).

Moreover, the data analysis allowed us to capture the distribution of social workers experience in the past (total experience) and compare it with the time of the research (to be called current experience). As depicted in **Figure 3**, responses concerning total and current experience were relatively similar at a statistically significant level (P-value < 0.05; hypergeometric test). To further describe this issue, we note that the polynomial trendlines and R2 formulas regarding total and current experience follow an identical prediction pattern. The "current experience" trendline seems to gain value after 3–4 years of experience. If we set "3–4 years" as the trendlines' curve alternation milestone, then we can substantially argue that social workers tend to remain in the same working position for less than the milestone of 4–5 years. This observation can be further explained but falls outside the scope of this article.

To summarize the aforementioned findings, it is obvious that our initial hypothesis (H0) has been validated through the comparison of statistically significant observations, as thoroughly described within this Section. Having validated H0 in Section "4 Our approach," we present our alternative approach to testing H0 through a modified K-Modes algorithm.

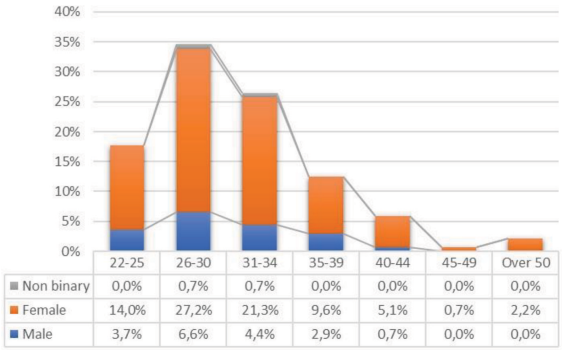

FIGURE 1 | Respondents' age and gender.

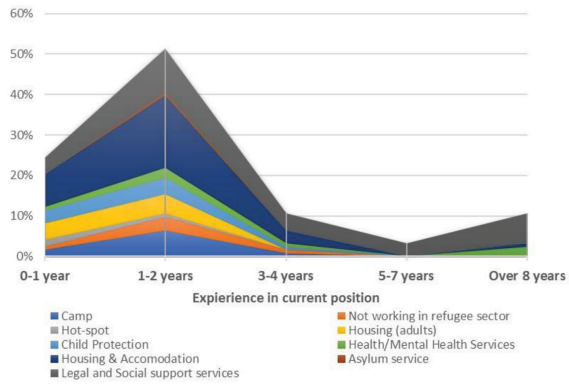

FIGURE 2 | Social workers' experience in current working position.

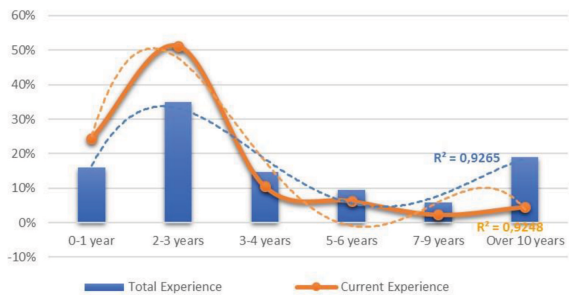

FIGURE 3 | Comparing social workers' total with current experience.

4. Our approach

As stated in Section "3 Social workers' profile: Validating our initial hypothesis," our main hypotheses concern the fact that social workers working in the front line of refugee and immigration services are, in the majority, young graduates with/or limited work experience (1). In order to validate the afore-mentioned hypothesis, in this section we propose a clustering algorithm in order to categorize data points into intuitively similar—but not predefined—user groups. This process is particularly useful toward summarizing the data and understanding the basic features that differentiate one user group (Cj) from another (44). In other words, our basic goal is to distinguish a set of users who possess the main characteristics of the attributes mentioned in Section "3 Social workers' profile: Validating our initial hypothesis."

4.1. Selecting input features for the clustering algorithm

Prior to clustering, our first task is to determine the input features. The importance stands on the fact that many research questions (features) present high diversity

regarding their answers (e.g., 90% of the social workers who have participated in the survey hold a master's degree). These features can create a common problem known as overfitting (45) and can misleadingly be considered the main features that define the differences in users' categorization. Plenty of methods can be utilized to define the optimal subset of features that have to be excluded from the clustering algorithm. Additionally, one of our major concerns during the feature selection process was the lack of an obvious way to evaluate the efficiency of feature selection without any specific domain knowledge capable of guiding the above process.

An indirect method of performing feature selection is to implement several scenarios with subsets of the available features and validate the efficiency of the clustering process for each scenario. This method may delay the desired features' extraction but guarantees high efficiency and reliability. Toward this, we decided to implement a greedy algorithm that creates all possible subsets while we evaluate the algorithm's efficiency for each subset.

As shown in **Figure 4**, every time a subset is generated, it undergoes through the clustering process and then gets evaluated. We define the subset with the minimum evaluation score as the optimal result of the "feature selection" phase. In order to calculate the evaluation score, a variety of metrics have been proposed in the literature, suggesting "intracluster to intercluster distance ratio" (46) being suggested as the most reliable. The idea behind this method is that the members of the same cluster should be "closer" to each other than the distance from other clusters' members. Consequently, the following metric is adopted, as defined in Aggarwal (44)

$$Intra = \frac{1}{|P|} \sum_{(I_i, I_j) \in P} dist(I_i\, I_j) \tag{1}$$

$$Inter = \frac{1}{|Q|} \sum_{(I_i, I_j) \in Q} dist(I_i\, I_j) \tag{2}$$

$$dist(I_i\, I_j) = \frac{1}{1 + Sim\left(I_i\, I_j\right)} \tag{3}$$

where $Sim\left(I_i, I_j\right)$ represents the overall similarity between two users. An above average clustering algorithm should produce results with a relatively small intra/inter ratio. Having calculated the ratio for every possible feature's subset, we select the subset with the minimum value.

4.2. Clustering with modified k-modes algorithm

Alternative clustering algorithms may apply to different data. In this occasion, we implement a modified k-Modes method

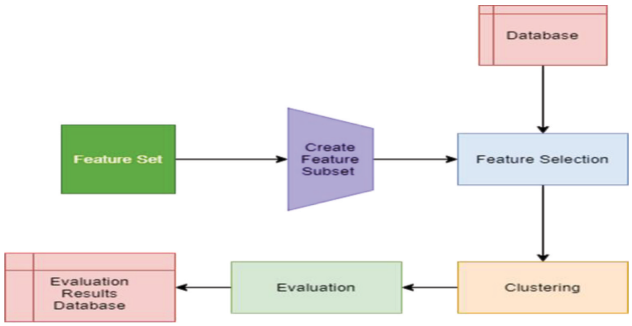

FIGURE 4 | Selecting the set of features for the clustering algorithm.

(6, 47) taking into consideration that most of our data contain categorical values. The k-Modes algorithm consists of the following steps:

(1) First, the number of clusters N_c to be created is chosen, and one representative for each group is randomly picked up from the initial database instances. The set of the representatives for all N_c will be referred as $T_R \in R^{N_c}$.

(2) In the second step, the similarity of each user to every representative R_j is calculated and inserted into a matrix $\in \mathbb{R}^{N_D \times N_c}$, where N_D is the size of our database (i.e., the number of social workers who participated in the research).

(3) Then, each user instance I_i is assigned to one of the N_c clusters, by determining the maximum of the similarity metrics for the row s_i of S, where $s_i \in \mathbb{R}^{N_c}$, $i = 1, 2, , N_D$. The cluster participation vector $C \in \mathbb{R}^{N_D}$ (i.e., to which cluster every user belongs) is then straightforwardly implemented by the following formula:

$$c_i = index_j(\max_j (s_{ij}))for\, j = 1, 2, , N_c and \quad \forall\, i = 1, 2, N_D \tag{4}$$

(4) Having assigned each user I_i to a N_c cluster, the new representatives R_j must be calculated. The *k-Modes* algorithm suggests that, for every cluster, a "virtual" user is created who does not belong to the original database. Therefore, we observe the cluster members for every feature F_r, and we calculate the frequency $p_r(x_r^l)$ of the x_r^l values. The value that is most common among the members is assigned to the new representative R_j. This method is referred to as taking the "mode" of the cluster members (6). If more than one of the x_r^l values are the most frequent for a feature F_r, the mode randomly selects one of them.

(5) Finally, we iterate over steps 2 to 4 until either the set of the representatives ($T_R^{old} = T_R^{new}$) or

the cluster participation vector ($C^{old} = C^{new}$) remain the same.

At this point, we should note that the *k-Modes* parameters are tuned appropriately to achieve optimal accuracy and efficiency in final clusters. Generally speaking, the parameters that need to be determined are (a) the number of clusters N_c, (b) the similarity metric used to categorize the social workers, and (c) the features used for the clustering process (also referred to as feature selection). Considering our main hypothesis that most of the social workers are young graduates and/or of limited experience, the value $N_c = 2$ is chosen. Explaining further this concept, we assume that if two distinguished groups of users are obtained, the largest of them containing most social workers who meet the standards of our hypothesis, then our hypothesis is indeed proven.

Regarding the metric used to calculate the similarity between users, we consider the following users: $I_i = \left(x_1^i, x_2^i, x_{N_F}^i\right)$ and $I_j = (x_1^j, x_2^j, x_{N_F}^j)$, where x_r^i and x_r^j are the values of the r^{th} feature for users $I_i$and I_j, respectively, while N_F represents the total number of features used to describe a user. The overall similarity between two users is then defined as follows:

$$Sim\left(I_i\, I_j\right) = \sum_{r=1}^{N_F} S(x_r^i, x_r^j) \tag{5}$$

where $S\left(x_r^i, x_r^j\right)$ denotes the similarity measure for the individual attribute value x_r of the two users. The most straightforward technique is to set $S\left(x_r^i, x_r^j\right) = 1$, for every feature value that x_r^i, x_r^j share in common. However, such a metric does not imply to the overall data distribution, meaning that the same similarity value is assigned in case where either the value appears in most of the data or it is of rare frequency. To overcome this drawback, we use the, as referred in the literature, "inverse occurrence frequency" metric (48). This metric calculates the similarity between matching attributes of two users by leveraging the weight of "rare" attribute values as follows:

$$S\left(x_r^i\, x_r^j\right) = \begin{cases} \frac{1}{p_r(x_r^i)^2}, if\ x_r^i = x_r^j \\ 0, otherwise \end{cases} \tag{6}$$

where $p_r(x) = \frac{N_r(x)}{N}$, and $N_r(x)$ stands for the number of users that possess the value x for the rth feature. One problem we encountered with similarity metrics was that in case of two features and specifically in the attributes "*Current Organization Type*" and "*Past Organization Type*" of our dataset, most of the users had multiple selections in a single answer. A workaround for this problem was to perform a modified one-hot encoding of these features. One-hot encoding (49) is the procedure of producing binary features labeled as all the possible values an initial feature can take. However, this can lead to a common problem in data science knows as the "curse of dimensionality" (50). To overcome this problem, we applied the following technique:

The term X_r^i refers to the set of values for the r^{th} feature of the user I_i takes. Accordingly, the term X_r^j refers to the set of values for the r^{th} feature of the user I_j. $X_r^{i\cap j}$ will be the set that is produced from the intersection of X_r^i and X_r^j:

$$X_r^{i\cap j} = X_r^i \cap X_r^j \tag{7}$$

The similarity $S\left(X_r^i, X_r^j\right)$ of the users for the r^{th} attribute can be defined as the sum of the similarity scores for the individual values that belong to the $X_r^{i\cap j}$ set. The above concept is shown in Equation 8.

$$S\left(X_r^i X_r^j\right) = \sum^{|} X_r^{i\cap j}|_{l=1} \frac{1}{p_r\left(x_l^{i\cap j}\right)^2} \tag{8}$$

where $x_l^{i\cap j}$ is the lth value in the $X_r^{i\cap j}$ set and $p_r(x) = \frac{N_r(x)}{N_D}$, where $N_r(x)$ stands for the number of users that possess the value x for the r_{th} feature.

In general, this technique bears great resemblance with the one-hot encoding with difference that the algorithm does not have to deal with the "curse of dimensionality," as the creation of new virtual features is avoided with the *ad-hoc* computation of the intersection set $X_r^{i\cap j}$ and the "inverse occurrence frequency" measure for the set's elements.

The overall concept of the afore-mentioned technique is presented as pseudo-code as follows:

```
begin:
Initialize the set of representatives T_R
with random users from database.
Initialize the similarity matrix S with
zeros.
while (T_R^old ! = T_R^new) do begin:
    for every user I_i:
      for every representative R_j:
        s_ij = 0;
        for every feature F_r:
         if F_r contains multiple value:
         for every value x_r in F_r that I_i
and R_j have in common:
            s_ij = 1/p_r(x_r)^2
         else if value x_r of F_r is the
same for I_i and R_j :
            s_ij = 1/p_r(x_r)^2
        Determine the cluster of I_i by
calculating c_i = index_j(max_j(s_ij))
        Append c_i to the cluster
participation vector C
```

```
    if (C^new == == C^old):
        break;
    for every cluster C_j:
        Determine the new representative
R_j and append it to T_R^new
return Cluster Participation Vector C
```

Algorithm 1 | *Modified k-Modes* (Database D, Number of Clusters N_C).

4.3. Evaluation of the clustering algorithm performance

Our dataset consists of 136 entries and 12 features. Two of the features ("*Current Organization Type*" and "*Past Organization Type*") can contain multiple values. To evaluate our results, we compare our clustering method with (a) the normal k-Modes algorithm (one-hot encoding for the features that contain multiple values) and (b) the agglomerative clustering method (51). In **Table 1**, we present the best three subsets of features for each technique as they were generated by the feature selection phase, as well as the *intra*/*inter* score for each subset. It is obvious that the overfitting that one-hot encoding causes does not lead to qualitative clustering results, as it can add too much "noise" to the features. As a result, the modified k-Modes we propose manage to reduce the *intra*/*inter* ratio values rather than the other clustering techniques.

To further examine the behavior of the three algorithms, the curve of the mean *intra*/*inter* ratio to the number of features used was designed (**Figure 5**). This allows us to observe and understand what to expect from these techniques in regards to how many features we are willing to discard. It is proven that the modified k-Modes outperform the other two clustering methods. In addition, while the mean *intra*/*inter* ratio is rising for hierarchical clustering and normal k-Modes as the number of features increases, the case is not the same for modified k-Modes. Not only does the mean ratio decrease, but as can be seen, the mean does not change drastically if we increase the number of features. This analysis indicates that the ideal number of features to be chosen is 5, as this value offers the advantage of fewer candidate features while maintaining appropriate levels of clustering efficiency. This assumption is additionally supported by the fact that the optimal subset produced through the feature selection method for the Modified k-Modes algorithm indeed had five features.

Taking the subset of features that yield the best evaluation score for the modified k-Modes algorithm T_F = {*Age, hasMSc, Total Experience, Current Organization Type, Total Time in the Current Organization*} leads us to the creation of two clusters C_1 and C_2. The first cluster contains 86 members, and the second contains 51 users. The scatterplot regarding "age" and "total experience" is designed in order to compare the distribution of these two clusters. The centroids (representatives) of each cluster are also presented for informational purposes in **Figure 6**.

Due to the categorical nature of these features, the following problem arose: if two users had the same values for the features age and total experience, the latter would cover the former. The solution to this problem was reached by adding Gaussian noise to every point on the plot so that the points would slightly differentiate around the initial spot. It is obvious that the members of the cluster C_1 are in their majority social workers that are young people (22–30 years old) and that their working experience is very limited, as

TABLE 1 | Algorithms' performance evaluation results.

Algorithm	Feature subset	Ratio
Modified k-Modes	*Age, hasMSc, Total experience, Current organization type, and Total time in the current organization*	**0.3653**
	Age, Total experience, Current organization type, Position at current organization, and Total yime in the past organization	**0.3719**
	Age, Total experience, Total time in the current organization, Position at current organization, and Total time in the past organization	**0.3733**
Agglomerative clustering	*Sex, Education, and hasMSc*	**0.5656**
	Education, hasMSc, and Position at current organization	**0.5837**
	Education, hasMSc, and Location of residence	**0.6156**
Normal k-Modes	*Age, Total experience, Total time in the current organization, Position at current organization, and Total time in the past organization*	**0.3733**
	Age, Location of residence, Total experience, Total time in the current organization, Position at current organization, and Total time in the past organization	**0.3852**
	Total Experience, Total time in the current organization, Position at current organization, and Total time in the past organization	**0.4027**

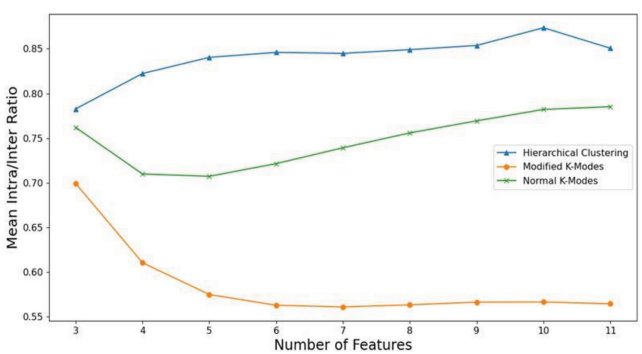

FIGURE 5 | Mean ratio to number of features for all algorithms.

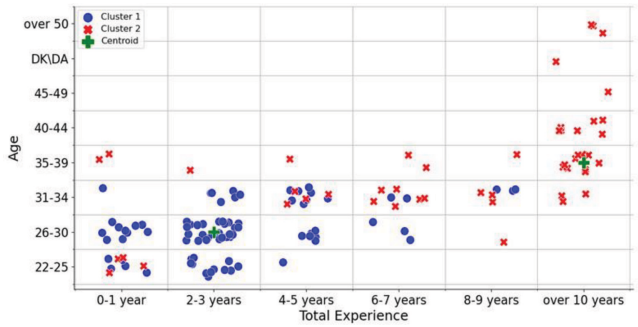

FIGURE 6 | Creating clusters with age/total experience scatterplot.

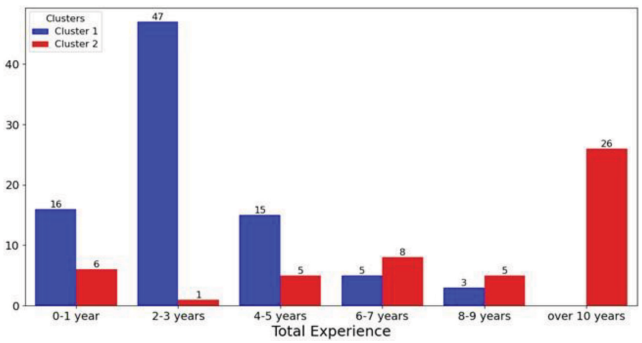

FIGURE 7 | Total experience distribution in clusters C1 and C2.

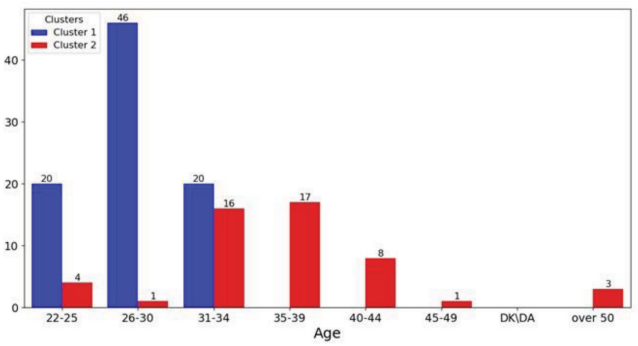

FIGURE 8 | Age distribution in clusters C_1 and C_2.

only 8 of them have worked for more than 5 years. On the contrary, the users that were categorized in C_2, have more than 10 years of experience and their age distribution is mainly 35 years old and above.

To further analyze our results, we created two graphs that compare the "age" and "total experience" distribution between the two clusters. As shown in **Figure 7**, in the cluster C_1 47 social workers have 2–3 years of experience, while 16 of them have 0–1 year of experience. This indicates that most of the social workers in this cluster have limited working experience.

Finally, observing the age distribution for the same cluster in **Figure 8**, it is straightforward that these social workers are also young graduates, as most of them are aged less than 30 years.

5. Conclusion

Social sciences cross the milestone of the artificial intelligence era in a rapidly changing scientific sub-field such as machine learning with enhanced data analysis tools as well as improved algorithms and techniques. Statistical hypotheses and testing through a variable significance level have been the norm for a very long period, demanding a "statistical" point of view into various problems and datasets.

Contrary to the above concept, we proposed an innovative approach that arises from clustering algorithms and aspires to become common ground for social science researchers in the upcoming years. Our approach exploits a modified k-Modes algorithm and bypasses statistical hypothesis testing through clustering construction. Moreover, we addressed the problem of selecting a subset of important features for the whole data in order to be aware of the "important" features before performing clustering. Consequently, the clustering process becomes more efficient, focused, and strict as only the important features are used. Therefore, our approach can be classified as a two-step method: we first rank and then select the subset of important features.

Finally, as a future work direction, it is important to note that the outcomes obtained through our approach, can be further evaluated with a number of alternative algorithms, and, in this regard, the researchers are free to apply their own technique and method selection or amendment and to reapply it if necessary. While clustering allows us to identify the sorting and allocation of observations, offering possibilities for researchers to study, we start with an initial number of clusters and then try to allocate the observations to correspondent clusters, with a future evaluation of the representativeness of each variable when creating them. Therefore, the result of one method can serve as input to the other, making this a "cyclical approach" or, as we define it, "***a recursive feedback method***".

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Teloni DD, Dedotsi S, Telonis AG. Refugee 'crisis' and social services in Greece: social workers' profile and working conditions. *Eur J Soc Work.* (2020) 23:1005–18.
2. Jones C. Voices from the front line: state social workers and New Labour. *Br J Soc Work.* (2001) 31:547–62.
3. Teloni DD, Dedotsi S, Lazanas A, Telonis A. Social work with refugees: examining social workers' role and practice in times of crisis in Greece. *Int Soc Work.* (2021) 1:1–18.

4. Lancaster HO, Seneta E. Chi-square distribution. *Encycl Biostat.* (2005) 2.
5. Adams WT, Skopek TR. Statistical test for the comparison of samples from mutational spectra. *J Mol Biol.* (1987) 194:391–96.
6. Bai L, Liang J. The k-modes type clustering plus between-cluster information for categorical data. *Neurocomputing.* (2014) 133:111–21.
7. Kurita T. An efficient agglomerative clustering algorithm using a heap. *Pattern Recogn.* (1991) 24:205–09.
8. Chaturvedi A, Green PE, Caroll JD. K-modes clustering. *J Class.* (2001) 18:35–55.
9. Zambom AZ, Collazos JA, Dias R. Functional data clustering via hypothesis testing k-means. *Comput Stat.* (2019) 34:527–49.
10. Ferraty F, Vieu P. *Nonparametric functional data analysis: theory and practice*. Berlin: Springer Science & Business Media (2006)
11. Horváth L, Kokoszka P. *Inference for functional data with applications*, Vol. 200. Berlin: Springer Science & Business Media (2012).
12. Hsing T, Eubank R. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Vol. 997. Hoboken, NJ: John Wiley & Sons (2015).
13. Kokoszka P, Reimherr M. *Introduction to functional data analysis*. Boca Raton, FL: Chapman and Hall/CRC (2017).
14. Jacques J, Preda C. Functional data clustering: a survey. *Adv Data Anal Class.* (2014) 8:231–55.
15. Wang JL, Chiou JM, Müller HG. Functional data analysis. *Annu Rev Stat Appl.* (2016) 3:257–95.
16. Reimherr M, Nicolae D. A functional data analysis approach for genetic association studies. *Ann Appl Stat.* (2014)8:406–29.
17. Young DL, Fields S. The role of functional data in interpreting the effects of genetic variation. *Mol Biol Cell.* (2015) 26:3904–08.
18. Bowman FD, Guo Y, Derado G. Statistical approaches to functional neuroimaging data. *Neuroimaging Clin N Am.* (2007) 17:441–58.
19. Hasenstab K, Scheffler A, Telesca D, Sugar CA, Jeste S, DiStefano C. et al. A multi-dimensional functional principal components analysis of EEG data. *Biometrics.* (2017) 73:999–1009.
20. Di Salvo F, Ruggieri M, Plaia A. Functional principal component analysis for multivariate multidimensional environmental data. *Environ Ecol Stat.* (2015) 22:739–57.
21. Saeys W, De Ketelaere B, Darius P. Potential applications of functional data analysis in chemometrics. *J Chemometr.* (2008) 22:335–44.
22. Aguilera AM, Escabias M, Valderrama MJ., Aguilera-Morillo MC. Functional analysis of chemometric data. *Open J Stat.* (2013) 3:334.
23. Eisen MB. Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA.* (1998) 95:14863–8.
24. Dahl DB, Newton MA. Multiple hypothesis testing by clustering treatment effects. *J Am Stat Assoc.* (2007) 102:517–26.
25. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat Sci.* (2003) 18:71–103.
26. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Ser B.* (2004) 66:187–205.
27. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat.* (2001) 29:1165–88.
28. Ward JH Jr. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc.* (1963) 58:236–44.
29. Hartigan JA. *Clustering algorithms*. Hoboken, NJ: John Wiley & Sons, Inc (1975).
30. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* (2002) 97:611–31.
31. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics.* (2002) 18:1194–206.
32. Finch H. Comparison of distance measures in cluster analysis with dichotomous data. *J Data Sci.* (2005) 3:85–100.
33. Tarpey T, Kinateder KK. Clustering Functional Data. *J Class.* (2003) 20.
34. Yamamoto M. Clustering of functional data in a low-dimensional subspace. *Adv Data Anal Class.* (2012) 6:219–47.
35. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol Methods.* (2012) 17:228.
36. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Computat Stat Data Anal.* (2014) 71:52–78.
37. Sakamoto Y, Ishiguro M, Kitagawa G. *Akaike information criterion statistics*. Dordrecht: Kluwer Academic Publishers Group (1986).
38. Bouveyron C, Jacques J. Model-based clustering of time series in group-specific functional subspaces. *Adv Data Anal Class.* (2011) 5:281–300.
39. Chiou JM, Li PL. Functional clustering and identifying substructures of longitudinal data. *J R Stati Soc Ser B.* (2007) 69:679–99.
40. Sugar CA, James GM. Finding the number of clusters in a dataset: an information-theoretic approach. *J Am Stat Assoc.* (2003) 98:750–63.
41. Giacofci M, Lambert-Lacroix S, Marot G, Picard F. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics.* (2013) 69:31–40.
42. Ciollaro M, Genovese CR, Wang D. Nonparametric clustering of functional data using pseudo-densities. *Electron J Stat.* (2016) 10:2922–72.
43. Bongiorno EG, Goia A. Classification methods for Hilbert data based on surrogate density. *Comput Stat Data Anal.* (2016) 99:204–22.
44. Aggarwal CC. *Data mining: the textbook*. Berlin: Springer (2015).
45. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Comput Surv.* (1995) 27:326–7.
46. Ray S., Turi RH. "Determination of number of clusters in k-means clustering and application in colour image segmentation," in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, New Delhi: Narosa Publishing House (1999), 137–43.
47. He Z, Xu X, Deng S. Attribute value weighting in k-modes clustering. *Expert Syst Appl.* (2011) 38:15365–9.
48. Havrlant L, Kreinovich V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *Int J Gen Syst.* (2017) 46:27–36.
49. Dahouda MK, Joe I. A Deep-learned embedding technique for categorical features encoding. *IEEE Access.* (2021) 9:114381–91.
50. Rust J. Using randomization to break the curse of dimensionality. *Econometrica.* (1997) 65:487–516.
51. Müllner D. fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. *J Stat Softw.* (2013) 53:1–18.