Available Online at http://www.journalajst.com

ISSN: 0976-3376

ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Asian Journal of Science and Technology
Vol. 16, Issue, 01, pp. 13356-13361, January, 2025

RESEARCH ARTICLE

SIMSYMB: A SIMILARITY MEASURE FOR HETEROGENEOUS DATA IN THE ANALYSIS OF SOCIAL RESILIENCE PROCESSES

COULIBALY Kpinna Tiekoura*, Kamagaté Beman Hamidja, Kanga Koffi and Soro Etienne

LASTIC, African Higher School of Information and Communication Technologies (ESATIC), Côte d'Ivoire

ARTICLE INFO

Article History:
Received 11th November, 2024
Received in revised form
20th December, 2024
Accepted 28th December, 2024
Published online 17th January, 2025

Keywords:

Social resilience measure of similarity, distance, automatic classification.

ABSTRACT

The assessment of resilience is the core of any governance and risk management analysis for shocks and crises. This assessment is generally dependent on similarity measurement for the identification of patterns and relationships between different individuals or groups of individuals within a given community. One of the difficulties in studying social resilience processes is the lack of appropriate analytical tools that take into account the dimensions of resilience. The use of conventional similarity measures can lead to some bias in the analysis and consequently to errors in decision-making. In this paper, we propose a new measure of similarity for the calculation of the degree of similarity between two individuals described by several univalued and multivalued variables of heterogeneous types (quantitative, qualitative or textual). Our proposal, compared to most of the similarity measures presented in the literature, has the merit of directly exploiting a table of heterogeneous data containing both univalued and multivalued values (intervals, sets, textual, etc.). Generally, a homogeneous transformation of the table is used and then a classical similarity index is used for the construction of the similarity matrix. However, this homogeneity of the table leads to distortions and negatively influences the expressive character of the data. The comparison of our approach with other proposals in the literature according to the Davies-Bouldin and Silhouette quality index gives us the best values for these indices, demonstrating its effectiveness for studying social resilience processes.

Citation: COULIBALY Kpinna Tiekoura, Kamagaté Beman Hamidja, Kanga Koffi and Soro Etienne. 2025. "Simsymb: A Similarity measure for Heterogeneous data in the Analysis of Social Resilience Processes", *Asian Journal of Science and Technology*, 16, (01), 13356-13361.

INTRODUCTION

In an increasingly interconnected world, social resilience is emerging as a key polysemic concept for understanding how individuals and communities adapt and recover in the face of crisis and disruption. However, the diversity of available data, both quantitative and qualitative, poses a major challenge for the analysis and comparison of resilience processes. Heterogeneous data, from a variety of sources such as surveys, community reports, government statistics or individual testimonies, require innovative approaches to extract meaningful insights. However, much of the existing work on resilience can be summed up in a broad transdisciplinary panel of non-operational theoretical approaches. As a result, the study of social resilience processes is hampered by the lack of appropriate modeling and analysis tools. Indeed, one of the most important aspects in the operational study of social resilience is the grouping of traumatized individuals into psycho-sociological strata, in order to optimize decision-making in their care. Furthermore, the automatic classification methods used in their application, for the construction of different classes, measures of similarity which are metric or semi-metric used in several fields including machine learning and data mining. However, literature is full of a multitude of measures of similarity that are usually quite general and do not take into account the specificities of resilience. Thus, there is a need to develop measures of similarities more adapted to the nature of data of social resilience processes. In this research work, we propose a new measure for calculating the degree of similarity between traumatized individuals or communities from heterogeneous data (quantitative, qualitative or textual). This measure will enable practitioners of resilience process analysis to obtain more robust and refined typologies of traumatized individuals, without having to make a first homogeneous transformation of the data table for the construction of the similarity matrix. Our approach will first present some related work on similarity measures while examining the issues related to data diversity in social resilience analysis. Next, we will present our methodological approach, detailing the construction of the similarity measure and its theoretical foundations. Finally, we will illustrate the application of this measure through an experiment on a data set, in order to demonstrate its effectiveness and relevance in the evaluation of social impact processes.

State of the art : In the literature, most of the similarity measures encountered are most often adapted to univalued-described data where each observation is represented by a point in space. These data tables may contain only data of a quantitative or qualitative type or even mixed. In all these cases, the formalism of the data used has a limited expressive power because it does not take into account complex and heterogeneous data from real applications as in the analysis of social transformation processes. In practice, observations can be described by symbolic variables, that is, textual values, sets of values, intervals and many other forms. These symbolic variables can be of the same type or different types (mixed symbolic variables). In the case where all variables involved are quantitative univalued, Minkowski distances [1] are generally used to measure similarity between observations. These distances, including the Euclidean distance and the distance from Manhattan, are a generalization of distances in a vector space, and they are often used in the context of geometry and number theory. For very small (but positive) values of the parameter p which determines the "standard" used, the Minkowski distance can become very sensitive to small variations in coordinates, which influences the quality of the results. When all data are interval type, the comparison of observations can be made using distance based on Hausdorff distance [2]. This distance has several limits, including its sensitivity to isolated or outliers points and its non-sensitivity to the internal structure. Thus two sets can have the same distance from Hausdorff while having very different shapes. When observations are described by symbolic variables of different types, data homogenization techniques are most often used to transform the table of heterogeneous data into a table of homogeneous data containing variables of the same type [3][4]. However, this technique usually leads to distortion and loss of information in the results, hence the need to develop similarity measures more suited to this type of data table. In the literature, the dissimilarity measure proposed by Gowda and Diday [5] and that of Ichino and Yaguchi [6] attempt to solve this problem. However, these measures are less suited to text-based data. Indeed, they do not take into account certain important aspects of textual similarity, such as the frequency of terms in the proximity calculus. The use of Gowda and Diday distances is often limited in practical applications for large datasets, where calculating distances can become costly in terms of computation time and resources, which can limit their use.

Ichino and Yaguchi's comparison function, on the other hand, is a generalization of the Minkowski distance based on a new mathematical model they call the Cartesian space model (U(p), ⊞, ⊠) where U(p) is the p-dimensional space of variables of different types, ⊞ a Cartesian joint operator and ⊠ a Cartesian intersection operator. In addition to the above-mentioned limitation, Ichino and Yaguchi's com-parison function uses a parameter γ to control the internal and external closeness between two observation intervals. An inappropriate choice of this parameter would distort the analysis, because if the intervals are disjoint, for example, the comparison function only takes into account the external closeness, ignoring the internal closeness of the intervals. The similarity measure (phi-similarity) proposed by Achiepo and Behou [7] allows the comparison of observations described by univalent quantitative, qualitative and textual variables, but is not adapted to mixed symbolic data. For their part, Stéphanie et al [8] have developed a similarity measure adapted from Gower's general similarity coefficient [9], known as the vulnerability similarity coefficient (HVSI). This method quantifies vulnerability profiles with the aim of identifying places with similar vulnerability, in order to facilitate the construction of networks for disaster resilience. This approach also aims to facilitate the sharing of knowledge, resources and successful practices that are relevant to the circumstances of a particular community. One of the weaknesses of this approach is that it does not always take into account the context in which a vulnerability arises, such as the operating environment, specific configurations or safety measures in place, and its sensitivity to false positives and false negatives, which can affect confidence in the results. Finally, at the level of unstructured data, Reimers et al [10] use cosine similarity to comparse semantically significant sentence embeddings, through their SBERT approach. While this considerably reduces the time needed to find the most similar pair, while maintaining good accuracy, SBERT is costly in terms of computational resources and requires adjustments for specific tasks as well as fine-tuning for specific domains.

Problematic: In a context where contemporary societies are increasingly confronted with multiple and varied crises, understanding social resilience processes becomes a crucial issue to anticipate and manage these challenges. However, the analysis of social resilience with tools that do not take into account its specificities often leads to bias in results and therefore to errors in decision-making. In other words, the measures of similarity, which are essential in models for grouping or stratifying individuals, must be adapted to the analysis of resilience processes and also to the ability to exploit various types of data. The presentation of the main measures of similarity in literature has allowed us to highlight several limitations, including that relating to the exploitation of heterogeneous data which can come from various sources such as surveys, demographic data, economic indicators, social networks, etc. In this context, it is important to ask the question of how to develop a measure of similarity, Able to integrate these heterogeneous data to provide relevant and nuanced analysis of social resilience processes? This issue raises methodological and theoretical issues, particularly in terms of processing data relating to resilience, modelling and interpretation of results, while seeking to meet the growing need for analytical tools adapted to the complex realities of social resilience processes.

Our proposal: SymSimb

Modeling : To take into account mixed univalent and symbolic data, as well as textual data, we combine the generalized Minkowski distance [1] and the classical cosine distance [11]. The former is based on Ichino and Yaguchi's mathematical model of Cartesian space and is adapted to quantitative and qualitative data from univalent and symbolic tables. The cosine distance is used to compare variables with textual descriptions.

Basic concepts: The data used in our similarity measurement model are essentially univalent qualitative and quantitative data, interval data, multivalued nominal and ordinal data, modal data and textual data from a corpus. In the following, we will use the term “symbolic data” to designate the various types of data mentioned, with the exception of textual data.

Let $\Omega = \{X_i, \ldots, X_n\}$, the set of individuals or observations X_i compared.

Λ, the set of variable domains considered in the data table.

$\Lambda = \Lambda_{(S)} \cup \Lambda_{(T)}$ …………………………(E1)

where $\Lambda_{(S)}$ represents the set of symbolic variable domains used.

$\Lambda_{(S)} = \Lambda_{S1} \times \Lambda_{S2} \times \ldots \times \Lambda_{Sq}$ …………………………(E2)

$\Lambda_{(T)}$ represents the domain of textual variables used.

The measure of similarity between two individuals X_i and X_j is defined by:

$SimSymb(X_i, X_j) = \lambda \Phi_S(X_i, X_j) + (1 - \lambda)\Phi_T(X_i, X_j)$ ……..(E3)

With $\Phi_S(X_i, X_j)$, the function of measuring the proximity between individuals X_i and X_j according to the symbolic variables.

$\Phi_T(X_i, X_j)$, the function measuring the proximity between individuals X_i and X_j according to textual variables.

and λ, the parameter for adjusting the influence of text variables in relation to symbolic variables and vice versa.

The $\Phi_S(.)$ function for calculating the similarity between individuals according to symbolic variables:

This function is based on the Minkowski distance and uses the mathematical model proposed by Ichino and Yaguchi [6] to compare the values taken by the symbolic variable Y_k characterizing individuals X_i and X_j.

Note:

- Y_k, the kth symbolic variable characterizing individualX_i ;
- x_i^k, the value taken by the variable Y_k describing the individual X_i ;
- η, the parameter controlling internal and external matching between two value intervals.
- N_S, the number of symbolic variables considered.

Let be the Cartesian space $(\Lambda_{(S)}, \oplus, \otimes)$ where $\Lambda_{(S)}$ denotes the q-dimensional space of the different types of symbolic variables used;

denotes the Cartesian operator of joint union and $\bigotimes$, the Cartesian operator of intersection such that:

$$X_i \oplus X_j = \prod_{k=1}^{N_S} \left(x_i^k \oplus x_j^k\right) \qquad \text{………..(E4)}$$

$$X_i \otimes X_j = \prod_{k=1}^{N_S} \left(x_i^k \otimes x_j^k\right) \qquad \text{……..…..(E5)}$$

The function of comparison between the values taken by the symbolic variable Y_k characterizing the individuals X_i and X_j is thus defined:

$$\varphi(x_i^k, x_j^k) = \left(\left|x_i^k \oplus x_j^k\right| - \left|x_i^k \otimes x_j^k\right|\right) + \eta\left(2\left|x_i^k \otimes x_j^k\right| - \left|x_i^k\right| - \left|x_j^k\right|\right) \qquad \text{……(E6)}$$

We'll use the value recommended by Ichino for parameter η, i.e. 0.5. Thus, the φ function becomes:

$$\varphi(x_i^k, x_j^k) = \left|x_i^k \oplus x_j^k\right| - \frac{\left|x_i^k\right| + \left|x_j^k\right|}{2} \qquad \text{……(E7)}$$

The aggregation function for comparing two individuals characterized by symbolic variables is therefore given by:

$$\Phi_S(X_i, X_j) = \left(\sum_{k=1}^{N_S} \varphi(x_i^k, x_j^k)^p\right)^{1/p} \qquad \text{………(E8)}$$

Taking p = 2, we obtain the generalized Euclidean distance.

$$\Phi_S(X_i, X_j) = \sqrt{\sum_{k=1}^{N_S} \varphi(x_i^k, x_j^k)^2} \qquad \text{………..(E9)}$$

Finally, we obtain:

$$\Phi_S(X_i, X_j) = \sqrt{\sum_{k=1}^{N_S} \left(\left|x_i^k \oplus x_j^k\right| - \frac{\left|x_i^k\right| + \left|x_j^k\right|}{2}\right)^2} \qquad \text{…………(E10)}$$

The $\Phi_T(.)$ function for calculating the similarity between individuals according to textual variables:

Cosine similarity is one of the most widely used similarity measures for text data. It is generally used to measure the similarity between two documents. It involves calculating the cosine of the angle between the vector representations of the documents to be compared. In the field of social resilience, we can use this measure to compare two given individuals. For each individual, we match the corpus that characterizes him or her. This may be a speech or an opinion on a given subject. The full version of each corpus is first transformed into a vector of character strings, also known as a “bag of words”, which describes the content of the document. Words are independent and their order is not important. The resulting bag-of-words contains the relevant terms, after sup-pressing the full text, empty words and punctuation, and then proceeding with lemmatization and segmentation. Following this step, each word is assigned a weight, which can be obtained booleanly or by word frequency. In our model, we use word frequency, which is obtained by counting the number of occurrences of the term in the document.

According to our model, the values taken by a textual variable, in the data table, are word vectors, i.e. the relevant terms extracted from the corpora.

Notation :

- Y_k, he k^th textual variable characterizing the individual X_i ;
- x_i^k, he set of relevant terms taken by variable Y_k for individual X_i ;
- x_j^k, the set of relevant terms taken by variable Y_k for individual X_j ;
- $\Gamma_{ij} = x_i^k \cap x_j^k$, the set of terms common to X_i and X_j for variable Y_k;
- $card(\Gamma_{ij})$, the number of terms common to X_i and X_j for variable Y_k;
- tf_{qi}^k, the frequency of occurrence of the term q corresponding to the vector of words x_i^k in the corpus characterizing the individual X_i.
- N_T, the number of textual variables considered.

The $\Phi_T(.)$ function for calculating similarity with text variables is therefore:

$$\Phi_T(X_i, X_j) = \frac{\vec{x}_i^k . \vec{x}_j^k}{\left\|\vec{x}_i^k\right\| . \left\|\vec{x}_j^k\right\|}$$

Considering the weights of the various terms, we obtain:

$$\Phi_T(X_i, X_j) = \frac{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k . tf_{qj}^k}{\sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k} . \sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k}} \qquad \text{(E12)}$$

The $SimSymb(.)$similarity measure between two individuals described by symbolic and textual variables is therefore calculated as follows:

Considering equations E10 and E12, we obtain:

$$SimSymb(X_i, X_j) = \lambda \Phi_S(X_i, X_j) + (1 - \lambda)\Phi_T(X_i, X_j) \qquad \text{(E13)}$$

$$SimSymb(X_i, X_j) = \lambda \sqrt{\sum_{k=1}^{N_S} \left(\left|x_i^k \oplus x_j^k\right| - \frac{\left|x_i^k\right| + \left|x_j^k\right|}{2}\right)^2} + \qquad \text{(E14)}$$

$$(1 - \lambda) \frac{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k . tf_{qj}^k}{\sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k} . \sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ij})} tf_{qi}^k}}$$

The λ parameter is obtained experimentally. It is used to adjust the influence of the textual part of the variables in relation to the symbolic part, and vice versa.

$$\lambda = \begin{cases} 1 & if\ \Lambda_{(T)} = \varnothing \text{(absence of text variables)} \\ 0 & if\ \Lambda_{(S)} = \varnothing \text{(absence of symbolic variables)} \\ \lambda \in]0;1[& if\ \Lambda_{(T)} \neq \varnothing\ and\ \Lambda_{(S)} \neq \varnothing \end{cases}$$

In case $\Lambda_{(S)} \neq \emptyset$ and $\Lambda_{(T)} \neq \emptyset$, we recommend taking $\lambda = 1/4$.

In addition, our similarity measure checks the similarity properties namely symmetry, positivity and maximality.

Symmetry:

$$SimSymb(X_i, X_j) = \lambda \sqrt{\sum_{k=1}^{N_S} \left(\left|x_j^k \oplus x_i^k\right| - \frac{\left|x_j^k\right| + \left|x_i^k\right|}{2}\right)^2} + \qquad \text{(E15)}$$

$$(1 - \lambda) \frac{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ji})} tf_{qj}^k . tf_{qi}^k}{\sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ji})} tf_{qj}^k} . \sqrt{\sum_{k=1}^{N_T} \sum_{q=1}^{card(\Gamma_{ji})} tf_{qi}^k}}$$

$$SimSymb(X_i, X_j) = Sim_G(X_j, X_i) \qquad \text{(E16)}$$

Positivity:

$\forall X_i$ and $\forall X_j \in \Omega$,

$$\lambda\sqrt{\sum_{k=1}^{N_S}\left(\left|x_j^k \oplus x_i^k\right| - \frac{\left|x_i^k\right|+\left|x_i^k\right|}{2}\right)^2} = \lambda\sqrt{\sum_{k=1}^{N_S}\left(\left|x_i^k \oplus x_i^k\right|\right)^2} = \lambda\sqrt{\sum_{k=1}^{N_S} {x_i^k}^2} \quad \text{(E17)}$$

In the same way

$$(1-\lambda)\frac{\sum_{k=1}^{N_T}\sum_{q=1}^{card(\Gamma_{ji})} tf_{qj}^k . tf_{qi}^k}{\sqrt{\sum_{k=1}^{N_T}\sum_{q=1}^{card(\Gamma_{ji})} tf_{qj}^k} . \sqrt{\sum_{k=1}^{N_T}\sum_{q=1}^{card(\Gamma_{ji})} tf_{qi}^k}} \geq 0 \quad \text{(E18)}$$

Hence $SimSymb(X_i, X_j) \geq 0$ (E19)

Maximality:

$$SimSymb(X_i, X_i) = \lambda\Phi_S(X_i, X_i) + (1-\lambda)\Phi_T(X_i, X_i) \quad \text{(E20)}$$

But $\Phi_S(X_i, X_i) = \sqrt{\sum_{k=1}^{N_S}\left(\left|x_j^k \oplus x_i^k\right| - \frac{\left|x_i^k\right|+\left|x_i^k\right|}{2}\right)^2}$ (E21)

$$\sqrt{\sum_{k=1}^{N_S}\left(\left|x_j^k \oplus x_i^k\right| - \frac{\left|x_i^k\right|+\left|x_i^k\right|}{2}\right)^2} = \sqrt{\sum_{k=1}^{N_S}\left(\left|x_i^k \oplus x_i^k\right|\right)^2} = \sqrt{\sum_{k=1}^{N_S} {x_i^k}^2} \quad \text{(E22)}$$

$$\sqrt{\sum_{k=1}^{N_S} {x_i^k}^2} \geq \sqrt{\sum_{k=1}^{N_S}\left(\left|x_i^k \oplus x_j^k\right| - \frac{\left|x_i^k\right|+\left|x_j^k\right|}{2}\right)^2} \quad \text{(E23)}$$

$$\lambda\Phi_S(X_i, X_i) \geq \lambda\Phi_S(X_i, X_j) \quad \text{(E24)}$$

In the same way, $(1-\lambda)\Phi_T(X_i, X_i) = \frac{\vec{x}_i^k . \vec{x}_i^k}{\|\vec{x}_i^k\| . \|\vec{x}_i^k\|}$ (E25)

$$(1-\lambda)\Phi_T(X_i, X_i) = (1-\lambda)(x_i^k)^2 . cos(\vec{x}_i^k, \vec{x}_i^k) = (1-\lambda)(x_i^k)^2 \quad \text{(E26)}$$

$$(1-\lambda)(x_i^k)^2 \geq (1-\lambda)(x_i^k)^2\, cos(\vec{x}_i^k, \vec{x}_j^k) \quad \text{(E27)}$$

$$(1-\lambda)\Phi_T(X_i, X_i) \geq (1-\lambda)\Phi_T(X_i, X_j) \quad \text{(E28)}$$

Therefore, $\lambda\Phi_S(X_i, X_i) + (1-\lambda)\Phi_T(X_i, X_i) \geq \lambda\Phi_S(X_i, X_j) + (1-\lambda)\Phi_T(X_i, X_j)$ (E29)

Hence $SimSymb(X_i, X_i) \geq SimSymb(X_i, X_j)$ (E30)

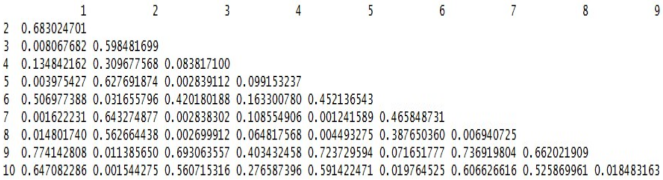
Algorithm: In machine learning, and particularly for geometric classification methods, similarity measures are used to construct similarity matrices between observations. The algorithm below shows the steps involved in calculating similarities between different observations in a table of heterogeneous data with symbolic and textual variables.

```
Algorithm: Calculating the similarity matrix for heterogeneous data (SimSymb)
Inputs : $\Omega = \{X_i, \ldots, X_n\}$, the set of individuals concerned;
         $\Lambda_S$, the set of symbolic variables used;
         $\Lambda_T$, the set of textual variables used;
         $\lambda$, the value of the parameter for fitting symbolic variables to textual variables;
         $\Delta_{tf}$, set of term frequencies in the corpus;

Outputs : M similarity matrix

Begin :

1. M [i, j] ← 0
2. For $i \leftarrow 1$ to n Do {
3.   while $i \leq j$ {
        3.1. Calculate $SimSymb(X_i, X_j) = \lambda\Phi_S(X_i, X_i) + (1-\lambda)\Phi_T(X_i, X_i)$
        3.2. M [i, j] = $SimSymb(X_i, X_j)$
        }
     }
4. Return M.
End
```

Experimentation : The advantage of our algorithm for measuring similarity, in contrast to most of the proposals presented in the literature, is its ability to directly exploit a table of heterogeneous data containing both univalent and multivalued values (intervals, sets, textual, etc.). Typically, a homogeneous transformation of the table is performed, followed by the construction of the similarity matrix using a classical similarity index. For example, qualitative data are transformed into quantitative data (and vice versa) and multivalued data are transformed into univalued data. However, this homogeneity of the array leads to distortions and negatively influences the expressive character of the data. In a field such as resilience, some survey data often express doubts between several values, or a range of values. Such information is then represented in the form of intervals or sets of values, in the data table. It is therefore important to keep the table as it is and use a generalizing similarity measure such as ours to construct the similarity matrix. To apply our measure, we use Table 1 of the data below. This table contains simulated data and illustrates the heterogeneity of the data. In this table, individuals are described by quantitative, qualitative and textual variables. Some variable values are univalent quantitative (e.g. Zadi's age) or univalent qualitative (e.g. Ahmed's marital status). We also distinguish between quantitative multivalued values (e.g. Yeo's and Kassi's age) and textual variables (e.g. individuals' opinions). The values taken by the textual variable (opinion) were obtained from different corpora, after elimination of punctuation, empty words and lemmatization. These different corpora correspond to the opinions of respondents to the question of the impact of the post-electoral crisis on their lives. The vector of terms or bag of words derived from the different opinions is as follows:

Γ = {crisis, election, impact, negative, positive, life, loss, loved one, forgiveness, not, repentance, violence, rape, sad, peace }

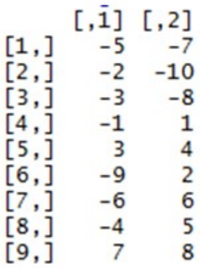
Implementing our algorithm in R, we obtain the following similarity matrix (Figure 1).



```
             1           2           3           4           5           6           7           8           9
2  0.683024701
3  0.008067682 0.598481699
4  0.134842162 0.309677568 0.083817100
5  0.003975427 0.627691874 0.002839112 0.099153237
6  0.506977388 0.031655796 0.420180188 0.163300780 0.452136543
7  0.001622231 0.643274877 0.002838302 0.108554906 0.001241589 0.465848731
8  0.014801740 0.562664438 0.002699912 0.064817568 0.004493275 0.387650360 0.006940725
9  0.774142808 0.011385650 0.693063557 0.403432458 0.723729594 0.071651777 0.736919804 0.662021909
10 0.647082286 0.001544275 0.560715316 0.276587396 0.591422471 0.019764525 0.606626616 0.525869961 0.01848316З
```

Figure 1. Similarity matrix (SimSymb)

This similarity matrix is used to group the 10 individuals into classes based on degrees of similarity. Figure 2 shows the different iterative groupings of individuals derived from this matrix.



```
     [,1] [,2]
[1,]   -5   -7
[2,]   -2  -10
[3,]   -3   -8
[4,]   -1    1
[5,]    3    4
[6,]   -9    2
[7,]   -6    6
[8,]   -4    5
[9,]    7    8
```

Figure 2. Distribution of individuals

According to Figure 2, over 9 iterations, the groupings are as follows:
Iteration 1 (C1) : ind 5 (Aïcha) & ind 7 (Armand)
Iteration 6 (C6) : ind 9 (Brou) & C2
Iteration 2 (C2) : ind 2 (Ahmed) & ind 10 (Koné)
Iteration 7 (C7) : ind 6 (Koffi) & C6
Iteration 3 (C3) : ind 3 (Paulette) & ind 8 (Kassi)
Iteration 8 (C8) : ind 4 (Yeo) & C5
Iteration 4 (C4) : ind 1 (Ahmed) & C1
Iteration 9 : C7 & ind C8
Iteration 5 (C5) : C3 & C4

Table 1. Symbolic data table

Names	Age	Monthly remuneration	Marital status	IQ	Gender	Opinion
Zadi	31	{400 000,450 000}	Married	9	M	{crisis, election, not, forgiveness, impact, negative}
Ahmed	25	[150 000,200 000]	Single	{8 ;9}	M	{crisis, election, forgiveness, impact, positive}
Paulette	{45,46}	170 000	Widow	{6 ;7 ;8}	F	{election, violence, rape, not, forgiveness}
Yeo	[60,90[	250 000	Married	6	M	{crisis, election, not, forgiveness, impact, negative}
Aïcha	29	75 000	Married	7	F	{crisis, forgiveness, impact, sad }
Koffi	63	1 000 000	Widow	{8 ;9}	M	{election, not, forgiveness }
Armand	28	120 000	Single	8	M	{crisis, election, Policy, impact}
Kassi	{32,33,34}	[200 000, 300 000]	Married	[6,9]	M	{crisis, election, not, forgiveness, sad}
Brou	85	{95 000, 150 000}	Single	9	F	{crisis, election, not, forgiveness, sad }
Koné	72	300 000	Marié	7	M	{election, forgiveness, impact, peace }

Table 2. DB and Silhouette Index values obtained with the 3 similarity measures

Number of clusters	IY		HVSI		SimSymb	
	DB	S	DB	S	DB	S
2 clusters	0,2901988	0,6781372	0,3576408	0,550101	0,3376408	0,7257529
3 clusters	0,2898222	0,5387805	0,3796834	0,4027573	0,2882724	0,6294764
4 clusters	0,313271	0,3888739	0,412069	0,3894272	0,3124571	0,4806501
5 clusters	0,201901	0,2673446	0,3925787	0,2948401	0,1966798	0,3506626
6 clusters	0,299423	0,290187	0,299423	0,3387971	0,2642	0,3433683
7 clusters	0,3690204	0,2682313	0,3730204	0,3113173	0,3590304	0,3253432

In order to evaluate our similarity measure (SimSymb) against existing heterogeneous data similarity measures in literature, we calculate the Davies-Bouldin index (DB) and the silhouette index (S) for each of the selected measures. These indices are used to assess the quality of the partitions formed from the distance matrices. The Davies-Bouldin index treats each class individually and seeks to measure how similar a class is to the class closest to it. The best partition, according to this index, is the one that minimizes the similarity between classes. In other words, the smaller the DB index, the better the partition. In addition, the Silhouette index is used to check whether each object has been correctly classified. The closer this index is to 1, the better the object is classified. For comparison, we use two similarity measures used in the construction of distance matrices for mixed data. These are the similarity measure of Ichino and Yaguchi [6] and that of Chang Stéphanie et al [8]. Table 2 below shows the DB and silhouette (S) indices corresponding to the similarity measures obtained with the 3 approaches. The graph in Figure 3 compares the three similarity measures according to the Davies-Bouldin quality index.

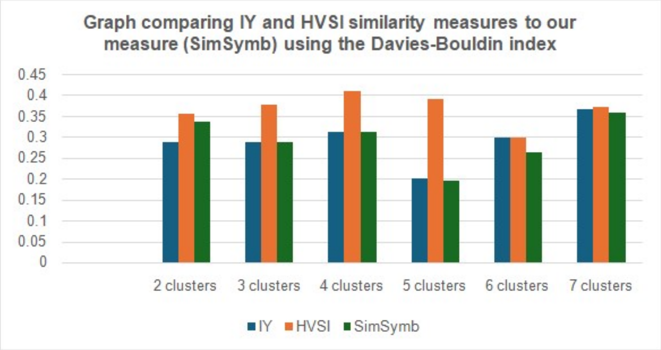


Figure 3. Comparison of similarity measures according to DB indices

Figure 3 shows the superiority of our similarity measure over the other two. Indeed, the smaller the DB index, the better the partition. Thus, partitions of 5, 6 and 7 clusters are of better quality with our measure than with the similarity of IY and HVSI. On the other hand, partitions containing two clusters are better with Ichino and Yaguchi similarity. For 3- and 4-cluster partitions, clustering quality is almost identical between our similarity measure and that of HVSI, with a slight superiority of our approach. The following graph (Figure 4) compares the three similarity measures according to the silhouette index. Beyond the quality of the clusters produced by these measures, it is also important to check, through the silhouette index, whether each individual is well classified using these similarity measures.

The silhouette index is widely used to assess the quality of clustering results. As a reminder, the silhouette score is a numerical value ranging from -1 to 1, which measures how well an observation is integrated into its group and how distinct it is from other groups. A value close to 1 indicates that the samples are well clustered, while a value close to -1 suggests that the samples may have been assigned to the wrong group.

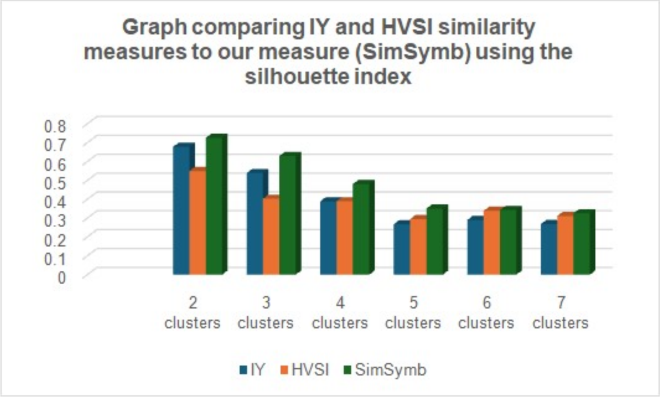


Figure 4. Comparison of similarity measures based on silhouette indices

Figure 4 also shows that our SimSymb similarity measure provides the best values for the silhouette index. This means that each individual is ranked higher with our similarity measure than with other measures.

CONCLUSION

In this article, we have presented a new measure of similarity between two given entities described by heterogeneous univalent or multivalued variables (numerical, qualitative and symbolic). Our model was validated using domain quality indices, notably the silhouette index and the Davies-Bouldin index. Our SimSymb proposal, compared to other measures, notably the similarity measure of Ichino and Yaguchi and the HVSI similarity measure of Chang Stéphanie et al. obtained the best values for these indices. This testifies to the quality of the groupings produced by our approach. By facilitating the integration and analysis of varied data, SimSymb contributes to a better understanding of resilience mechanisms, while offering new perspectives for researchers and practitioners engaged in the study of social dynamics. In order to assess the robustness of our proposal, it would be interesting in future work to test the algorithm on large-scale real data.

Acknowledgement: NÉANT.

REFERENCES

Achiepo, Odilon Yapo M., Behou Gérard N'Guessan, and Konan Marcellin Brou. "Similarity Measure in the Case-Based Reasoning Systems for Medical Diagnostics in Traditional Medicine." International Journal of Computer Science Issues (IJCSI) 12.2 (2015) : 239

Baeza-Yates R. et B. Ribeiro-Neto. Modern Information Retrieval. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass., 1999.

Chang, Stephanie E., et al. "Using vulnerability indicators to develop resilience networks: a similarity approach." Natural Hazards 78.3 (2015) : 1827-1841.

De Carvalho F.A.T. et R. M. C. R. De Souza: Unsupervised pattern recognition models for mixed feature-type symbolic data. Pattern Recognition Letters, 31:430–443, 2010

De Souza, R. M. C. R. F. A. T. De Carvalho et D.F. Pizzato: A partitioning method for mixed feature-type symbolic data using a squared Euclidean distance. In C. Freksa, M. Kohlhase et K. Schill, éditeurs : KI 2006 : Advances in Artificial Intelligence, volume 4314 de Lecture Notes in Computer Science, pages 260–273. Springer Berlin Heidelberg, 2007.

Gowda K.C. et E. Diday: Symbolic clustering using a new dissimilarity measure. Pattern Recognition Letters, 24(6):567–578, 1991

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4):857–871

Ichino M. et H. Yaguchi: Generalized minkowski metrics for mixed featuretype data analysis. IEEE Transactions on Systems, Man, and Cybernetics, 24 (4):698–708, 1994.

Ichino M. et H. Yaguchi: Generalized minkowski metrics for mixed featuretype data analysis. IEEE Transactions on Systems, Man, and Cybernetics, 24 (4):698–708, 1994

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D19-1410

Rote, Günter. "Computing the minimum Hausdorff distance between two-point sets on a line under translation." Information Processing Letters 38.3 (1991): 123-127

****