International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Taking into account Qualitative and Textual Variables in Hierarchical Ascending Clustering (HAC)

Odilon Yapo M. ACHIEPO, Kouassi Hilaire EDI, Behou Gérard N'GUESSAN, Patrice MENSAH

Abstract: ***In Machine Learning, the clustering methods are the mains unsupervised methods. Their objectives is to partition a set of objects in some homogeneously groups. Clustering methods in general and more particularly Hierarchical Ascending Clustering (HAC) techniques are based on metrics and ultra-metrics. Metrics are used to evaluate the similarities between two objects; and ultra-metrics are used to estimate the similarity of two groups or the similarity of an element and a group. The main characteristic of these metrics and ultra-metrics is the fact that they are only adapted to numerical variables or can be reduced to them. With the advent of Data Mining and Data Science, most of the datasets to be analyzed contain different types of variables. In the same dataset, we can find numeric attributes, qualitative variables and free text fields very often together. Despite this diversity of variables in the same dataset, the existed clustering methods are generally build to use only an unique kind of attribute. In this paper, we propose an approach to take account different types of attributes in the same clustering method. The method proposed is a variant of HAC methods that can take into account both numerical, qualitative and textual data. Our approach is based on a metric call Phi-Similarity we developed in order to estimate the proximity of two objects, each of them is describe by a vector of attributes of different types. The developed method will be implemented with the scientific computing language R and applied to real survey data. A comparison of the results will be made with HAC techniques based on classical metrics with the Ward criterion as aggregation criteria. For classical algorithms, we will limit ourselves to the variables of the database compatible with them. This work of comparison will highlight the gain in precision in terms of classification brought by our method compared to the classic versions of HAC***

Keywords : Hierarchical Ascending Clustering, Phi-similarity, R-Language

Manuscript published on 30 September 2019
* Correspondence Author
Odilon Yapo M. ACHIEPO*, University Péléforo Gon Coulibaly, Management Institut Agropastorale - Korhogo, Côte d'Ivoire. Email: kingodilon@gmail.com.
Kouassi Hilaire EDI, University Nangui Abrogoua, Mathematics and Computer Science Laboratory - Abidjan, Côte d'Ivoire. Email: edi.hilaire@gmail.com
Behou Gérard N'GUESSAN, Virtual University of Cote d'Ivoire, Research and Digital Expertise unit (UREN) Abidjan. Email: gerard.nguessan@gmail.com
Patrice Edoété MENSAH, National Polytechnic institute Felix Houphouet Boigny of Yamoussoukro Côte d'Ivoire, Email : pemensah@hotmail.com

I. INTRODUCTION

CLUSTERING is a very important area in Statistics and Artificial Intelligence, especially in the field of Machine learning and its applications. From a technical point of view, the clustering methods are of several types, including so-called hierarchical ascending clustering (HAC) methods. This category has the advantage not only of generating a visualization of the mechanism of grouping individuals in the form of a dendrogram but also, and above all, provide a possibility of automatically determining the optimal number of groups to perform. HAC methods are based on two fundamental criteria: a similarity index and an aggregation index ([1]). Regarding the measure of similarity between individuals to partition, many measures of "distance" have been proposed including the Euclidean distance, the distance from Manhattan, the distance from Camberra, etc. For aggregation indices, these are metrics that calculate the distance between an individual and a class or between two classes without having to recalculate those **between** the individuals making up the classes. There are a large number of which the best known are the distance of the minimum jump, the distance of the maximum jump and the distance of the inertia (criterion of Ward). The algorithm of the HAC is based on the initial situation where each individual of E constitutes a group all by itself. Then to group individuals according to their proximities until we obtain a situation in which all individuals form one and the same class.

II. ALGORITHMIC ASPECT OF THE HAC

The HAC is a purely algorithmic procedure based on a principle of proximity enhancement between individuals ([2]). For a more technical presentation of this algorithm, let's ask:

- Ω_I, the set of individuals described by a set Ω_X of variables

Algorithm 1: Algorithm of HAC methods

1) Entry : (Ω_I, d, μ_d)
2) Take each couple with $(I_i, I_j) = (C_i, C_j)$ with $(I_i, I_j) \in \Omega_I \times \Omega_I$
3) For each couple $(C_i, C_j) \in \mathcal{P}(\Omega_I) \times \mathcal{P}(\Omega_I)$ to do
   a) If $Card(C_i) = 1$ and $Card(C_i) = 1$ so
      Calculate $d(C_i, C_j)$
      If not
      Calculate $\mu_d(C_i, C_j)$
      End if
   b) Conserve the results $d(C_i, C_j)$ or $\mu_d(C_i, C_j)$ calculated
4) Group the couples (C_i, C_j) the closest in a new class C_k
5) Take back from 3) until you get a single class C_{final}
6) Output : The three of successive groupings C_g until C_{final}

- $\mathcal{P}(\Omega_I)$, all parts of Ω_I
- d, a distance between individuals defined on Ω_X
- μ_d, an aggregation index associated with the distance d

On the basis of these notations, the HAC algorithm can be described as follows:

However, it is easy to see that classical clustering methods require that individuals be described by essentially numerical variables. They cannot be used outside this frame; which is a very serious limit because, in the current databases, the presence of qualitative and especially textual fields becomes

Algorithm 1: Algorithm of HAC methods

1) Entry : (Ω_I, d, μ_d)
2) Take each couple with $(I_i, I_j) = (C_i, C_j)$ with $(I_i, I_j) \in \Omega_I \times \Omega_I$
3) For each couple $(C_i, C_j) \in \mathcal{P}(\Omega_I) \times \mathcal{P}(\Omega_I)$ to do
   a) If $Card(C_i) = 1$ and $Card(C_i) = 1$ so
      Calculate $d(C_i, C_j)$
      If not
      Calculate $\mu_d(C_i, C_j)$
      End if
   b) Conserve the results $d(C_i, C_j)$ or $\mu_d(C_i, C_j)$ calculated
4) Group the couples (C_i, C_j) the closest in a new class C_k
5) Take back from 3) until you get a single class C_{final}
6) Output : The three of successive groupings C_g until C_{final}

the norm. Therefore, this article proposes to extend HAC methods to take into account these various types of data. The main difficulty would be to find an appropriate metric for the use of different types of data. The main difficulty would be to find an appropriate metric for the use of different types of data. This limit finds its solution in the Phi-Distance that we propose.

III. PHI-SIMILARITY AND PHI-DISTANCE

Many existing methods for making typologies (k-means, k-meloid, hierarchical clustering, etc.) are based on metrics ([3]). However, in the literature, the similarity measures used do not take into account the textual data. In addition, the approaches used for qualitative data are most often related to simple counts of individuals that may not be statistically meaningful. To do this, it seemed sensible to propose a more appropriate index of similarity. That is why this article proposes a metric able to take into account these three types of data simultaneously.

Thus, the proposed similarity index is developed respecting the mathematical properties required for a similarity measure ([5]); in particular the symmetry and the nullity for an individual in relation to himself.

IV. BASIC MATHEMATICAL NOTATION

Consider the following mathematical notation:

- $\mathcal{S} = (X_1, \ldots, X_S) = (X_j)_{1 \leqslant j \leqslant S}$,
- the set of S criteria (variables) taken into account in the evaluation of the proximities between the different individuals
- $\mathcal{N}$ the set of N numerical criteria among all the criteria taken into account in the evaluation of the proximities between the different individuals
- $\mathcal{Q}$ the set of Q categorical criteria among all the criteria taken into account in the assessment of proximities between different individuals
- $\mathcal{T}$ the set of T textual criteria among all the criteria taken into account in the evaluation of the proximities between the different individuals

First of all, we can write that $\mathcal{S} = \mathcal{N} \cup \mathcal{Q} \cup \mathcal{T}$ And we have $S = N + Q + T$.

Then, if we note I_s and I_c two individuals described by the criteria considered, and if we note $Sim(.)$ measuring the similarity between these two individuals and Φ a rewrite of $Sim(.)$, we have :

$$\begin{aligned} &Sim(I_s, I_c) = Sim(\mathcal{S}_s, \mathcal{S}_c) \\ &= Sim(\{\mathcal{N}_s \cup \mathcal{Q}_s \cup \mathcal{T}_s\}, \{\mathcal{N}_c \cup \mathcal{Q}_c \cup \mathcal{T}_c\}) \\ &= \Phi(\{(\mathcal{N}_s, \mathcal{N}_c), (\mathcal{Q}_s, \mathcal{Q}_c), (\mathcal{T}_s, \mathcal{T}_c)\}) \\ &= \Psi(\alpha(\mathcal{N}_s, \mathcal{N}_c), \beta(\mathcal{Q}_s, \mathcal{Q}_c), \gamma(\mathcal{T}_s, \mathcal{T}_c)) \\ &= \Psi(\alpha(I_s, I_c), \beta(I_s, I_c), \gamma(I_s, I_c)) \end{aligned}$$

The functions α, β, γ designate distances between two vectors of respectively numerical, qualitative and textual variables. They apply to every uniform part of the vector describing an individual. As to the function Ψ, it means the final mathematical expression of the similarity measure (Phi-distance)

V. THE FUNCTION α BETWEEN TWO DIGITAL VECTORS

α being a distance between two digital vectors, it seems natural to define it using a Minkowski distance. Therefore, in general, we have:

$$\alpha(I_s, I_c) = \sqrt[p]{\sum_{X_k \in \mathcal{N}} (X_{ks} - X_{kc})^p}.$$

In practice, we can limit ourselves to $p = 2$ which corresponds to the Euclidean distance. Thus, we will take as distance between the numerical variables, the function defined by:

$$\alpha(I_s, I_c) = \sqrt{\sum_{X_k \in \mathcal{N}} (X_{ks} - X_{kc})^2}$$

VI. THE FUNCTION β BETWEEN TWO VECTORS OF QUALITATIVE VARIABLES

β Being a distance between two vectors of qualitative variables, it seems natural to consider the number of common modalities between individuals as a basis for constructing the metric. However, it must be taken into account that there are several qualitative variables. If we note $X_{k\omega}$ the modality of the variable X_k taken by the individual I_ω, the difference in modality between individuals I_s and I_c relatively to the variable X_k is given by:

$$\beta_k(I_s, I_c) = \begin{cases} 1 & \text{si } X_{ks} = X_{kc} \\ 0 & \text{si } X_{ks} \neq X_{kc} \end{cases}$$

The metric β is defined by:

$$\beta(I_s, I_c) = \frac{K - \sum_{k=1}^{K} \beta_k(I_s, I_c)}{1 + \sum_{k=1}^{K} \beta_k(I_s, I_c)}$$

VII. THE FUNCTION γ BETWEEN TWO TEXTUAL DATA VECTORS

The function γ is a distance between two vectors of textual variables. To define it, we will consider the distance called « cosine distance ». Since several textual variables are used, we propose to construct a cosine distance per variable then, to sum their sum weighted by the relative weight of each textual variable relative to the sequences of common words (to a corrective factor close).

$\forall X_k \in \mathcal{T}$, we note Δ_{ks} and Δ_{kc} the corpora of the textual variable $X_k$corresponding respectively to individuals I_s and I_c obtained from the same sequence of linguistic operations, in particular the same operation of segmentation of texts (tokenization). We recommend text lowercase operations, clearing blanks, deleting numeric values, removing punctuation, removing blank words, and word tokenization. Once the tokenization is done, the set of words of X_k common to individuals I_s and I_c is given by$\Delta_k = \Delta_{ks} \cap \Delta_{kc}$.

We note $\tau_{kj} \in \Delta_k$ a term (word) encountered term δ_j^{ks} time in Δ_{ks} and δ_j^{kc} time inΔ_{kc}. The number of times the term τ_{kj} is met in Δ_k is$\delta_{kj} = Min(\delta_j^{ks}, \delta_j^{kc})$. The cosine distance is défined by :

$$cosine_k(I_s, I_c) = \frac{<X_{ks}, X_{kc}>}{\|X_{ks}\|.\|X_{kc}\|} = \frac{\sum_{j=1}^{card(\Delta_k)} \delta_j^{ks}.\delta_j^{kc}}{\sqrt{\sum_{j=1}^{card(\Delta_k)} \delta_j^{ks}} \sqrt{\sum_{j=1}^{card(\Delta_k)} \delta_j^{kc}}}$$

Considering the weights$\lambda_k = \frac{1+Card(\Delta_k)}{\sum_{k=1}^{k=T}(1+Card(\Delta_k))}$,

we obtain:

$$\gamma(I_s, I_c) = \sum_{k=1}^{k=T} \lambda_k (1 - cosine_k(I_s, I_c))$$

VIII. THE FUNCTION Ψ OF GLOBAL SIMILARITY

The measures α, β and γ are distances. To deduce a measure of similarity, we will use a polynomial interpolation of Lagrange on the interval$[0, 1]$. Indeed, let's noted_α, d_β and d_γ standardized values of α, β and γ. So we have $d_v = 1 - \frac{1}{1+v}$ with $v \in \{\alpha, \beta, \gamma\}$ and posing$\mu = \frac{d_\alpha + d_\beta + d_\gamma}{3} = 1 - \frac{\frac{1}{1+\alpha} + \frac{1}{1+\beta} + \frac{1}{1+\gamma}}{3}$, we have by construction$\mu \in [0, 1]$.

We want to build the function Ψ that $\Psi(0) = 100$, $(i \in [1, n])$, $\Psi(\frac{1}{2}) = 50$, $\Psi(\frac{3}{4}) = 25$ and $\Psi(1) = 0$. So we have in the plan a set of 5 points $(x_i, y_i)_{1 \leqslant i \leqslant 5}$ which is the whole of $\{(0, 100), (\frac{1}{4}, 75), (\frac{1}{2}, 50), (\frac{3}{4}, 25), (1, 0)\}$. The Lagrange polynomial associated with each pair (x_i, y_i) is given by $l_i(x) = \prod_{i=0, j \neq i}^{j=5} \frac{x - x_j}{x_i - x_j}$, which allows to obtain:

$$\Psi(x) = \sum_{i=1}^{i=5} y_i l_i(x) = 100(1 - x)$$

If we conform to the defined notations, we can take as a measure of global similarity, the quantity defined by:

$$Sim(T_s, T_c) = \Psi(\mu(T_s, T_c))$$

The index of similarity as defined is decreasing with the level of dissimilarity of individuals, reflexive, symmetric and bounded between 0and 100. Its projection on the interval [0, 1] defines the Phi-distance.

Algorithm 2 : Algorithm for calculating Phi-similarity

1) Entrance : $(I_i, I_j) \in \Omega_I^2$
2) Calculate $\alpha(I_i, I_j)$
3) Calculate $\beta(I_i, I_j)$
4) Calculate $\gamma(I_i, I_j)$
5) Calculate $\Psi(I_i, I_j) = 100[1 - (\frac{1}{1+\alpha(I_i,I_j)} + \frac{1}{1+\beta(I_i,I_j)} + \frac{1}{1+\gamma(I_i,I_j)})/3]$
6) Exit : $\Psi(I_i, I_j)$

IX. THE ALGORITHM FOR CALCULATING THE PHI-SIMILARITY MATRIX

In most methods using similarity indices, the aim is to evaluate the similarities between several individuals, which amounts to calculating the distances between all the possible pairs of individuals considered in the form of a matrix of similarities. NotingΩ_I, the set of individuals considered, the algorithm for calculating the Phi-similarity matrix between several individuals is as follows:

X. THE MUTHAC ALGORITHM

The MuTHAC algorithm a particular version of the HAC algorithm. It uses Phi-distance and associated aggregation criteria to extend it to qualitative and textual variables. For a more technical presentation of this method, consider the following notations:

- Ω_I, The set of individuals described by a set of Ω_X variables

$\mathcal{P}(\Omega_I)$, All parts of Ω_I

- α, The function between two digital vectors used in Phi-similarity

Algorithm 2 : Algorithm for calculating Phi-similarity

1) Entrance : $(I_i, I_j) \in \Omega_I^2$
2) Calculate $\alpha(I_i, I_j)$
3) Calculate $\beta(I_i, I_j)$
4) Calculate $\gamma(I_i, I_j)$
5) Calculate $\Psi(I_i, I_j) = 100[1 - (\frac{1}{1+\alpha(I_i,I_j)} + \frac{1}{1+\beta(I_i,I_j)} + \frac{1}{1+\gamma(I_i,I_j)})/$
6) Exit : $\Psi(I_i, I_j)$

- β, The function between two qualitative vectors used in Phi-similarity
- γ, The function between two textual data used in Phi-similarity
- μ_α, an aggregation index associated with α
- μ_β, an aggregation index associated with β
- μ_γ, an aggregation index associated with γ

On the basis of these notation,
The MuTHAC algorithm is as follows:

Algorithm 3 : MuTHAC Algorithm

1) Entrance : $(\Omega_I, \alpha, \beta, \gamma, \mu_\alpha, \mu_\beta, \mu_\gamma)$
2) Take each couple $(I_i, I_j) = (C_i, C_j)$ with $(I_i, I_j) \in \Omega_I \times \Omega_I$
3) For each couple $(C_i, C_j) \in \mathcal{P}(\Omega_I) \times \mathcal{P}(\Omega_I)$ to do
   a) If $Card(C_i) = 1$ and $Card(C_i) = 1$ so
      i) Calculate $\alpha(C_i, C_j)$
      ii) Calculate $\beta(C_i, C_j)$
      iii) Calculate $\gamma(C_i, C_j)$
      iv) Calculate $\Psi(C_i, C_j) = 100[1 - (\frac{1}{1+\alpha(C_i,C_j)} + \frac{1}{1+\beta(C_i,C_j)} + \frac{1}{1+\gamma(C_i,C_j)})/3]$
   If not
      i) Calculate $\mu_\alpha(C_i, C_j)$
      ii) Calculate $\mu_\beta(C_i, C_j)$
      iii) Calculate $\mu_\gamma(C_i, C_j)$
      iv) Calculate $\mu_\Phi(C_i, C_j) = 100[1 - (\frac{1}{1+\mu_\alpha(C_i,C_j)} + \frac{1}{1+\mu_\beta(C_i,C_j)} + \frac{1}{1+\mu_\gamma(C_i,C_j)})/3]$
   Final if
   b) Keep the results $\Psi(C_i, C_j)$ or $\mu_\Phi(C_i, C_j)$ calculated
4) Group the couples (C_i, C_j) the closest in a new class C_k
5) Take back from 3) until you get a single class C_{final}
6) Exit : The tree of successive groupings C_g up to C_{final}

Table 1 : Description of database

N°	Variable	Nature	Description of variable
01	*woman*	Numerical	Number of women in the household
02	*man*	Numerical	Number of men in the household
03	*assistance*	Numerical	Amount of financial assistance received by the household
04	*education*	Numerical	Education expenditure by the household
05	*transport*	Numerical	Transportation expenditures by the household
06	*clothing*	Numerical	Clothing expenditure by the household
07	*health*	Numerical	Health expenditures made by the household
08	*social environment*	Qualitative	Living area of the respondent
09	*gender*	Qualitative	Gender of the respondent
10	*nationality*	Qualitative	Ivoirian status on not of the respondent
11	*spirituality*	Qualitative	Type de spirituality or religion
12	*Bank credit*	Qualitative	Opinion of the respondent on bank credit
13	*field*	Qualitative	Branch of activity of the respondent
14	*age*	Qualitative	Age range of the respondent
15	*level of education*	Qualitative	Level of education of the respondent
16	*strate*	Textual	Current regional division of Côte d'Ivoire
17	*former*	Textual	Former regional division of Côte d'Ivoire
18	*region*	Textual	Region of the respondent
19	*department*	Textual	Department of the respondent
20	*diploma*	Textual	Diploma obtained by the respondent
21	*stoping*	Textual	Reason that led the respondent to leave school
22	*function*	Textual	Current function of the respondent
23	*activity*	Textual	Type of activity exercised by the respondent
24	*status*	Textual	Statut de l'enquêté dans son activité professionnel
25	*place*	Textual	Workplace of the respondent
26	*origin*	Textual	Country of origin of the respondent

XI. PRETREATMENT OF VARIABLES

The database contains both classic variables (quantitative and qualitative) as well as text variables (open questions). The use of such a database with both classical variables (quantitative and qualitative) and textual variables in classical data analysis procedures (factor analysis, clustering) is impossible. Indeed, only quantitative and qualitative variables can be exploited in classical algorithms. However, such textual data are very common in large-scale statistical surveys. And although these data are collected, their level of exploitation is almost null. Hence the interest of our approach which consists in systematically taking into account the textual data collected at the individual level in the classical methods of clustering. To do this, the raw database is subjected to a preprocessing procedure. The purpose of this procedure is to label non-numeric variables with more than 10 different responses as textual and to handle any missing values. The preprocessing procedure used is described as follows:

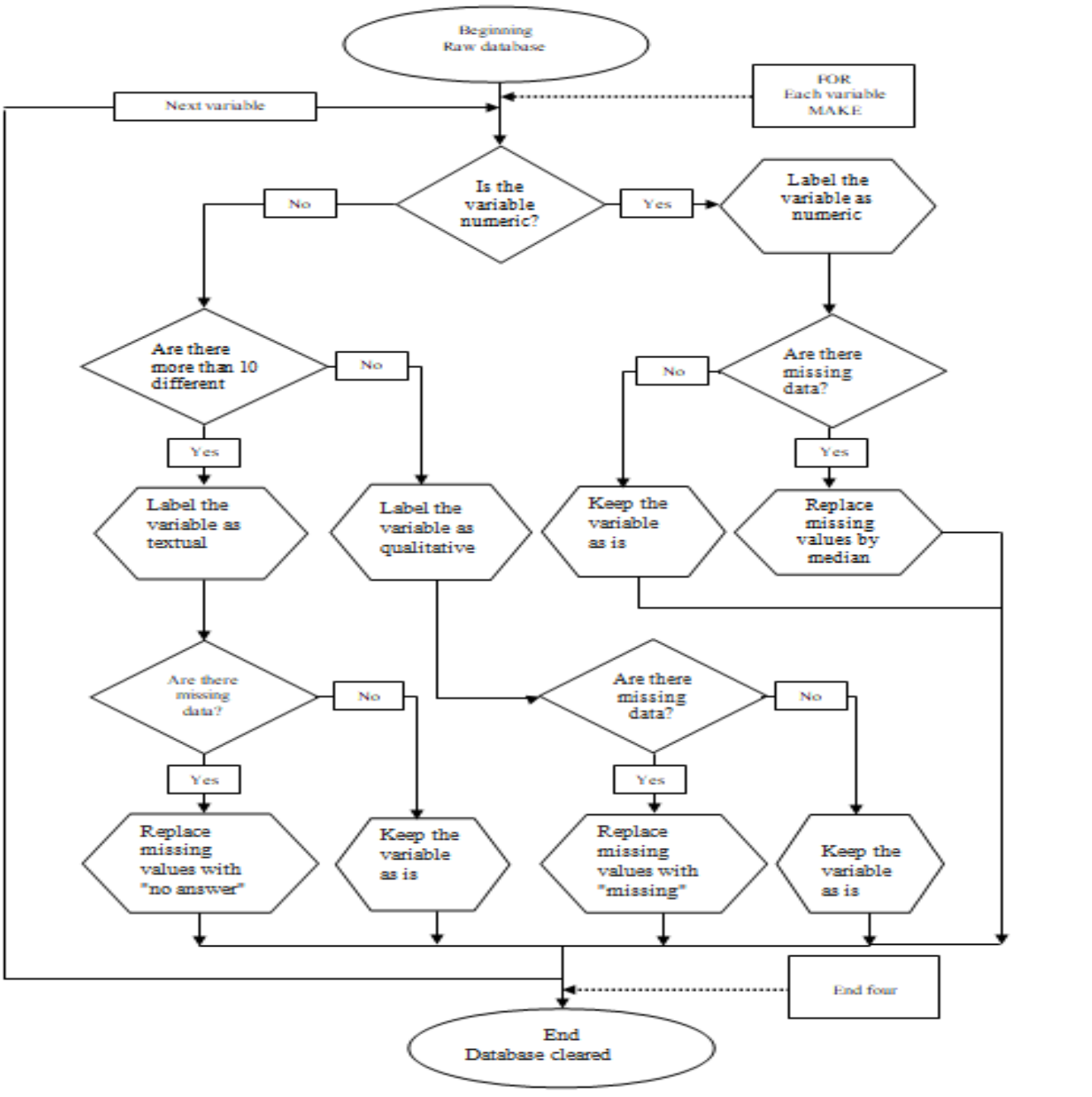

Fig 1 : Preprocessing variables flow chart

13. Application Results

All simulations were performed using the programming language R ([6], [8]). The results obtained with our method are compared with those obtained using classical metrics ([4)]. The performance of the clustering is measured with the Dunn index and connectivity. The results of the calculations are recorded in the following table :

Table 2 : Results simulation with R

Performance Indicator	Metric	Nomber of clusters				Nomber of optimal clusters
		K=2	K=3	K=4	K=5	
Dunn Index	PhiDist	0,61*	0,64*	0,63*	0,66*	K=2
	Euclidean	0,05	0,02	0,03	0,01	K=5
	Maximum	0,05	0,08	0,02	0,01	K=5
	Manhattan	0,02	0,02	0,03	0,00	K=5
	Canberra	0,25	0,25	0,25	0,23	K=5
Connectivity (5 first neighbors)	PhiDist	30,08*	34,60*	42,15	50,97	K=5
	Euclidean	7,00	19,77	26,00	40,80	K=5
	Maximum	6,36	12,44	30,95	42,59	K=5
	Manhattan	11,98	19,12	25,48	36,79	K=5
	Canberra	16,52	28,78	43,18*	64,00*	K=5

*High Value

The results show that, whatever the metric used, the best score is that in 5 classes. The Dunn index is a measure of homogeneity of partitions, the higher it is, the better the score is. The results of the simulations show that our metric, Phi-distance, always produces the best partitions in terms of homogeneity, compared to the majority of the classical distances used. As far as connectivity is concerned, it measures the encroachment of classes on each other. The result obtained with Phi-distance is mixed, especially with the distance from Canberra. If we stick to the best partition, namely class partitioning, we see that the Phi distance produces a level of connectivity that is not as good as the Euclidean distance, the distance of the maximum jump and the distance from Manhattan; but better than the distance from Canberra. This metric is therefore halfway between high connectivity and low connectivity methods. Thus, taking into account the qualitative and textual variables makes it possible to obtain a compromise in terms of connectivity which results in a very high level of performance in terms of homogeneity of the partitions with respect to all the classical distances that do not exist operate only on quantitative variables.

XII. Conclusion

This article proposes an clustering method that takes into account quantitative, qualitative and textual variables. This algorithm is a particular version of the ascending hierarchical clustering method. The main interest of this article is to extend the possibilities of clustering analysis on the many non-uniform attributes datasets we found in the research, industry and professional services. The simultaneous taking into account variables of different types is made through the definition of a metric called Phi-distance. This metric is general and can be use in many methods out of the clustering domain. The developed method is use on a and simulations are carried out with the R programming language. The real data used is from the Ivory Coast Living Standards Survey in 2015, yielded satisfactory results. These calculations have shown that Phi-distance we proposed, always produces the best partitions in terms of homogeneity, compared to the majority of the classical distances used. This exceptional performance measured by the Dunn index is obtained at the cost of less good connectivity than the scores obtained with the Euclidean distance, the distance of the maximum jump and the distance from Manhattan, however the degree of connectivity obtained by the Phi-distance remains better than that obtained with the distance of Canberra.

REFERENCES

1. Antoine Cornuejols, Laurent Miclet. “Apprentissage Artificiel, Concept et Algorithme” 2ème édition, Eyrolles, 2010.
2. Benzecri J.-P., Lebeaux M.-O., Jambu M. “Aides à l’interprétation en classification automatique”. Les Cahiers d’Analyse des Données, 5, p 101-123, 1983.
3. Eric Biernat, Michel Lutz. “Data Science: fondamentaux et etudes de cas. Machine learning avec Python et R”. Eyrolles 2016.
4. Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta. “Valid: An R Package for Cluster Validation”. Journal of Statistical Software, 25(4), 1-22, 2008.
5. I. Witten, E. Frank. (1999). Data Mining: Pratical Machine Learning Tools and Techniques. Morgan Kaufmann.
6. Jones O, Maillardet R, Robinson A. “Introduction to Scientific Programming and Simulation Using R”. Chapman & Hall/CRC, Boca Raton, 2009.
7. L. Miclet, S. Bayoudh, A. Delhay. “Analogical dissimilarity: Definition, algorithms and two experiments in machine leraning”. Journal of Artificial Intelligence Research., 32: 793-824, 2008.
8. R. Ihaka and R. Gentleman. “R: A language for data analysis and graphics”. Journal of Computational and Graphical Statistics, 5:299–314, 1996.

AUTHORS PROFILE

Odilon Yapo M. ACHIEPO is a Doctor in Computer Science. He holds a degree in Statistician-Economist Engineer at ENSEA in Abidjan (Côte d'Ivoire) and a Master's degree in Computer Science with a specialization in Knowledge Extraction from Data obtained at Polytechnic School of Nante's University (France). He is also a consultant, Teacher-researcher at Péléforo Gon Coulibaly University of Korhogo (Ivory Coast). His areas of expertise and interests include Statistical Engineering, Artificial Intelligence, Big Data, Internet of Things, Data Science, Fast Data and Ethical Hacking. He is also a member of the Laboratory of Data Engineering and Artificial Intelligence of Abidjan, and associate member of the Unit of Research and Digital Expertise of the Virtual University of Cote d'Ivoire. He is also the author-creator of Resiliometrics, a modeling discipline that consists of developing and applying computational models for measurement, analysis and simulation of social resilience process.

Kouassi Hilaire EDI is a Doctor in Industrial Systems Engineering. He has the PhD degree from the Department of Industrial Engineering at Toulouse University, ENSIACET, France. His research interests on the modeling of industrial application problems and artificial intelligence. He is working with other researchers to develop a data engineering model. Currently he works as assistant professor at Nangui Abrogoua University, Abidjan (Ivory Coast). At the same time, he is responsible for the MIAGE Master (Applied Computer Methods for Business Management) of the UFR of Fundamental and Applied Sciences (SFA). He is a member of the Laboratory of Mathematics and Computer Science of UFR-SFA and Associate Member of UREN (Unit of Research and Digital Expertise) of the Virtual University of Ivorydi Coast.

Behou Gérard N'GUESSAN is a doctor in computer engineering. He holds a Master's Degree in Media Engineering at 08-May-1945 university of Guelma, Algeria. He received his PhD at Nangui Abrogoua University, Abidjan-Côte d'Ivoire at the Faculty of Applied Basic Sciences. He is a member of the Research Laboratory in Computer Science and Telecommunications at the Houphouet Boigny National Polytechnic Institute (INP-HB), Abidjan, Côte d'Ivoire, member of Laboratory of Data Engeering and Artificiel Intelligence and associate member of the Unit of Research and Digital Expertise of the Virtual University of Cote d'Ivoire. His research interests include mathematical modeling, media engineering, traditional medicine and the inventor of applications. His work focuses on their research and training method in traditional medicine. He is currently working as an assistant professor at the Virtual University of Ivory Coast in Abidjan (Ivory Coast).

Patrice Edoété MENSAH holds a PhD in Applied Mathematics from St Louis University, Washington, USA. He is Associate Professor at the Department of Mathematics and Computer Science at Houphouet Boigny National Polytechnic Institute of Yamoussoukro (Ivory Coast). He is also Associate Professor at the International Institute of Water and Environmental Engineering of Ouagadougou (Burkina Faso). He is a member of the Laboratory of Data Engineering and Artificial Intelligence of Abidjan. He is also a member of the Research Mixt Unit and Innovation in Mathematics and New Information Technologies of INP-HB Yamoussoukro. His areas of interest include Applied Mathematics, Industrial Engineering and Control of Linear and Nonlinear Systems..