MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220185

Interpreting the Human Longevity Profile Through Triadic Rules - A Case Study Based on the ELSA-UK Longitudinal Study

Marta D. M. Noronha[a], Cristiane N. Nobre[a], Mark A. J. Song[a], Luis E. Zárate[a]

[a]Department of Computer Science, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

Human aging is a complex process with several factors interacting. One of the ways to identify patterns about human aging is longitudinal population studies. In this work, we identified longevity profiles through a process of knowledge discovery. After identifying the profiles, we apply triadic rules which allow extracting rules of implication with conditions. These rules can be used to identify related factors, in the various waves, of longitudinal studies, which can better explain the conditions that favor longevity profiles.The results show that the triadic analysis is efficient to allow the analysis of the temporal evolution of clinical or environmental conditions that favor certain profiles when databases of longitudinal studies are considered.

Keywords:

Aging, Public Health, Data Mining.

Introduction

Human aging is a complex process, with several genetic and environmental factors interacting, which are responsible for the changes that occur in people's lives as they age. Recently, the demand for knowledge about aging has increased, due to the increase in the elderly population in the world [1]. With a higher proportion of elderly people in society, interest in finding patterns of healthy habits increases in order to create public policies and social programs that increase the well-being of this portion of the population [2]. One of the initiatives to identify patterns about human aging is longitudinal population studies and data mining.

Longitudinal studies have followed a fixed set of people over several years. During these years, data related to a domain of study are collected, repeatedly, in fixed periods of time called waves. The databases generated by these studies are called longitudinal databases.

Traditionally, longitudinal data for aging studies are analyzed using classical statistical techniques, such as linear and non-linear regression, hypothesis testing, etc. [3]. In addition to these studies being focused on smaller sets of variables, statistical studies are often parametric, that is, they make assumptions about the distribution of data. The use of computational techniques based on machine learning allows a holistic analysis of the databases through non-parametric techniques, allowing the discovery of complex patterns present in the data.

The objective of this work is the application of a knowledge discovery process based on the set of data generated from the population study ELSA-UK (English Longitudinal Study of Aging) [16], using machine learning techniques, to identify profiles of short-lived and long-lived individuals.

After a pre-processing of the database, composed of examples of long-lived and short-lived ELSA participants, the bi-clustering technique was applied using the BiMax algorithm. This technique allowed to identify subsets of individuals that best represent the longevity profiles considering a subset of attributes. Based on the identified subsets, we propose the application of the theory of triadic concepts, derived from the theory of Formal Concept Analysis [4] to understand the evolution of conditions that favor longevity profiles.

The Formal Concepts Analysis (FCA) theory [4] is an applied mathematical field whose main objective is to represent and extract knowledge. In the last years, the FCA has received many attention in the Data Mining field [5] as a powerful technique to represent and extract knowledge from datasets expressed as cross tables, namely formal contexts.

The Triadic Concept Analysis (TCA) is a extension of the FCA theory, which allows extracting rules of implication with conditions, which can be used to identify related factors, in the various waves, of longitudinal studies, which can better explain the conditions that favor longevity profiles.

In general, describing longevity profiles can be significant for understanding how aging is influenced by environmental, socioeconomic and health factors.

The results show that the triadic analysis is efficient to allow the analysis of the temporal evolution of clinical or environmental conditions that favor certain profiles, when databases of longitudinal studies are considered.

Background

The FCA theory has some main definitions: Formal context, Formal concept and Implications rules. Formal contexts have the notation K:=(G, M, I), where G is a set of objects, M is a set of attributes and I is an incidence relation ($I \subseteq G \times M$). If an object $g \in G$ and an attribute $m \in M$ are in the relation I, this is represented by $(g, m) \in I$ or gIm and is read as "the object g has the attribute m".

Given a set of objects $A \subseteq G$ from a formal context K:=(G, M, I), it could be asked which attributes from M are common to all those objects. Similarly, it could be asked: "for a set $B \subseteq M$, which objects have the attributes from B". These questions define the derivation operators, which are formally defined as:

$$A' := \{m \in M \mid gIm \ \forall g \in A\} \quad (1)$$

$$B' := \{g \in G \mid gIm \ \forall m \in B\} \quad (2)$$

A conditional proposition P and Q is an implication P→Q, such that P and Q are sets of attributes and P'⊆Q' (see Equations 1 and 2). In other words, every object that has the attributes of P also has the attributes of Q. Note that, the implications P→Q can be extracted from a formal context K:=(G, M, I).

The TCA theory was introduced in [6]. Formally, a triadic context is defined as a quadruple (K1,K2,K3,Y), where K1, K2 and K3 are sets and Y is a ternary relation between K1, K2 and K3, i.e., Y ⊆ K1 x K2 x K3, the elements of K1, K2, and K3 are called (formal) objects, attributes, and conditions, respectively, and (g,m,b) ∈ Y is read: "the object g has the attribute m under the condition b". An example of a triadic context is represented in Table 1. The attributes correspond to 3 waves of the longitudinal study and the conditions to the symptoms persisting throughout the clinical follow-up s1… s4.

Table 1 – Example of triadic context

Att	*Wave 1*				*Wave 2*				*Wave 3*			
Obj	*s1*	*s2*	*s3*	*s4*	*s1*	*s2*	*s3*	*s4*	*s1*	*s2*	*s3*	*s4*
1	*X*	*X*		*X*	*X*	*X*		*X*	*X*		*X*	
2	*X*			*X*		*X*	*X*	*X*	*X*	*X*		*X*
3	*X*	*X*		*X*				*X*	*X*	*X*		
4	*X*	*X*		*X*		*X*		*X*	*X*	*X*		
5	*X*			*X*	*X*			*X*	*X*	*X*		*X*

According to the literature, [7] was the first work to deal with the problem to extract implication rules from a triadic context. After that, diferent kind of implications were developed by [8], such as Attributes x Condition Implications (AxCIs), Conditional Attribute Implications (CAIs). and Attributional Condition Implications (ACIs).

We describe two types of triadic association rules that can be extracted from a triadic context K= (K1,K2,K3,Y) by combining ideas [7], [8]. The two types considered in this work are: Biedermann Conditional Attribute Association Rule (BCAAR) and Biedermann Attributional Condition Association Rule (BACAR).

- BCAAR has the form*: (A1→A2)C(sup, conf),* where *A1, A2⊆K2* and *C ⊆K3* [9]. Its meaning is as follows: whenever *A1* occurs under all conditions in *C*, then *A2* also occurs with support *(sup)* and confidence *(conf)* as defined by [9].
- BACAR has the form*: (C1→C2)A(sup, conf),* where *C1, C2⊆K3* and *A⊆K2* [9]. Its meaning is as follows: whenever *C1* occurs for all attributes in *A* then the condition in *C2*, also occurs for the same attributes.

Methodology

Materials - ELSA-UK Longitudinal Study

ELSA is currently one of the most prominent population aging studies in the world [12]. The study has thousands of respondents (all inhabitants of the United Kingdom) aged 50 and over, interviewed every two years (duration of a study wave) by professionals for data collection. ELSA started in 2002, and its database contains demographic, economic, social, physical health, mental and psychological health, and cognitive function variables.

In order to identify representative instances for the profiles, the first 6 waves of ELSA were considered. Individuals who exceeded the UK's life expectancy (82.9 years), or died before reaching that age were labeled as long-lived and short-lived individuals, respectively.

Note that only one instance was maintained for each individual in this database, to avoid redundancy (in a longitudinal database, there are repeated instances referring to each individual). The last available record was maintained for short-lived individuals and the first available record for long-lived individuals. This was done to approximate the average age of individuals in both classes, in order to reduce the influence of age on the profiles, since the objective is to find longevity profiles in relation to the attributes observed in ELSA.

Method

Database pre-processing

The database then went through the following pre-processing steps (described in detail in [10]):

- Elimination of inconsistent instances or attributes, or considered unreliable.
- Application of the missing data technique Last-Observation-Carried-Forward [11]. Attributes with missing data were replaced by values of the same attribute, for the same instance, in an earlier study wave, when available.
- Conceptual selection of attributes, guided by a previous study to identify the environmental aspects related to human aging [12]. All attributes maintained in the database describe aspects used in other studies of human aging.
- A merger of related attributes was carried out to reduce the dimensionality of the database. Questions from the ELSA questionnaires that were directly dependent on each other were merged into a single attribute that represents all the information obtained in the set of dependent questions.
- All attributes (dichotomous, categorical or numerical) had their values transformed to a numerical value between 0 and 1. This recoding was done so that the database could be used as input to the machine learning algorithms, and follows a logic internal in the attributes, where all the smaller values are less desirable for a long life than the larger ones. That is, all attributes have a value of 0 for the "worst case" option, and a value of 1 for the "best case" option.
- Factor analysis was applied to the dichotomous variables (consequently, transforming the dichotomous variables into numeric ones), for the subsequent application of biclustering, as a semi-supervised technique, in the search for factors that may characterize the profiles (search by sub-groups). Short-lived labels were assigned to non-accidental deaths that occurred before an individual reached longevity. Similarly, for people who have attained longevity, longevity labels are assigned when the person reaches 82.9 years of life. Instances related to people where there is no information about their premature death have not been labeled, therefore being removed from the study presented in this work.

- Finally, after identifying factors and variables that best characterize the longevity profiles, the next step was to characterize one of the longevity profiles, namely, long-lived or short-lived through triadic analysis (due to space restrictions, in this work, only the longitudinal profile will be considered) where the instances that are present in the considered waves are analyzed using triadic association rules BCAAR and BACAR. For this work we consider the computational tool Lattice Miner 2.0 [15].

After identifying the most representative profiles, following the methodological procedures described in this section, the next step was to identify the triadic relationships between aspects (attributes in the triadic context) that influence longevity and are present along the waves (triadic conditions). For this, only the first 3 waves were used to form the triadic context due to the tracking of long-lived individuals being present in these waves.

For this work, 522 instances of long-lived profiles were selected from among the initial 1,091 identified in the approach based on biclustering. After a new coding to enable the formation of the longitudinal triadic context, similar to Table 1, only 50 instances of long-lived profiles were identified up to the third wave. It is important to note that triadic contexts require a great deal of computational effort to be processed, which may limit their applicability in large contexts.

Results

After the procedures, described above, have been applied to the ELSA-UK longitudinal base, we now describe the result of the biclustering process, through which it was possible to identify variables that contribute to characterize the longevity profiles.

Figures 1 and 2 show the factors (obtained by Factor Analysis) that characterize the subgroups of the profiles. Notice that, the Factor (F5) is a relevant discriminant to separate the two profiles.

Figure 1 – Average of the factors

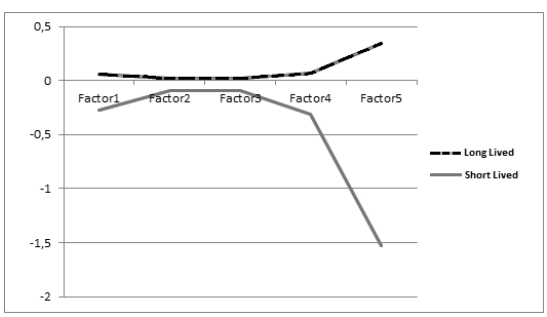
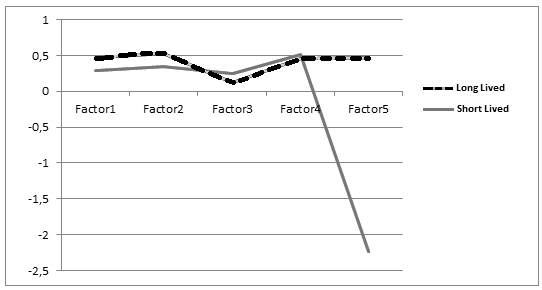


Figure 2 – Median of the factors



The variables represented by Factor F5 are described in Table 2. The variables: dhsex, hecat and iacisa_npb_sava (more details about the variables in [16]) were eliminated because they have a low correlation, bellow 0.30. After analysis, it was decided to separate the variables that make up "hoevm-hopay" and "iadebt-iaowe", being that "hoevm", "iaowe", "hevsi" and "wpbus" were discarded because they have a very low entropy.

Additionally, as recommended in [10], were considered variables related to social and emotional aspects, Table 3. The answers given to the questions by ELSA participants are scored in the biclustering approach as 0, 0.33, 0.66 and 1.0. But for use in TCA a dichotomization was made, ignoring the weights, to enable the formation of the triadic context. Therefore answers valued as 0 and 0.33 (for answers: "Frequently" or "Sometimes") are coded as zero (0), while 0.66 and 1.0 are coded as one (1) to favor the aspect of longevity (for answers: "Rarely" or "Never"). The same procedure was adopted for the variables in Table 2. For example, if the person does not inject insulin, a value of 1 is assigned, otherwise it receives a value of 0 and is not filled in the triadic context.

After the identification of these variables, we proceeded to the analysis by means of TCA to discover longitudinal rules that characterize the profile of long-lived. The Triadic context contains: 50 objects (long-lived individuals), 13 attributes (variables, Table 2 and 3); and 3 conditions (waves 1, 2, 3 of ELSA-UK).

Table 2 - Dichotomous features represented by relevant factor

Feature	Loading	Description
iadebt/iaowe	0.75	Reported having debt
hevsi	0.67	Reported eyesight difficulties due to stroke
hefrac	0.52	Reported hip fracture
hoevm/hopay	0.51	House owner
wpbus	-0.51	Business owner
heins	0.45	Injects insulin (diabetes)
mmhss	0.34	Walking test not taken due to health condition
hecanb	0.31	Cancer treatment in the last 2 years
wpedc	-0.31	Formal education in the last 12 months
*dhsex**	0.26	Gender (Male or Female)
*hecat**	-0.06	Had cataracts surgery
*iacisa_npb_sava**	0.26	Reported having savings

* Feature considered without significant for having loading lower than 0.3 ([13], [14]); 15).

Table 3–Aditional variables: Negative feeling

Feature	Weight	Description
scqola	3	Age prevents you from doing things you would like to
scqolb	5	What happens is out of your control
scqold	4	You are left out by others
scqolf	3	Family responsabilities prevent you from doing things you want to do
scqolh	5	You health stops you from doing things you want to do
scqoli	1	Lack of money stops you from doing things you want to do

After applying the TCA, using the Lattice Miner 2.0 tool, [15], on this longitudinal context 52 rules of the BACARs type and 1,968 rules of the BCAARs type were obtained, with support and confidence above 50% each. Among these we show the rules with support above 95%. It is important to note that triadic rules are rules of association explored within the context. Hence it is possible to obtain a high number of rules. For further analysis, it is possible to adjust the support and confidence measure to identify the most relevant ones.

BCCARS

1: (hefrac→ mmhss) w2 [sup = 100.0% conf = 100%]

2: (hefrac→ iadebt) w2 [sup = 96.0% conf = 96.0%]

3: (mmhss→ iadebt) w2 [sup = 96.0% conf = 96.0%]

4: (mmhss→ hefrac) w2 [sup = 100.0% conf = 100%]

5: (hefrac→ iadebt) w3 [sup = 98.0% conf = 98.0%]

6: (hefrac→ mmhss) w3 [sup = 96.0% conf = 96.0%]

7: (heins→ mmhss) w1 [sup = 96.0% conf = 98.0%]

8: (heins→ hecanb) w1 [sup = 96.0% conf = 98.0%]

9: (iadebt→ hefrac) w3 [sup = 98.0% conf = 100%]

10: (iadebt→ hefrac,mmhss) w2 [sup = 96.0% conf = 100%]

11: (mmhss→ heins) w1 [sup = 96.0% conf = 98.0%]

12: (mmhss→ hecanb) w1 [sup = 96.0% conf = 98.0%]

13: (hecanb→ heins) w1 [sup = 96.0% conf = 98.0%]

14: (hecanb→ mmhss) w1 [sup = 96.0% conf = 98.0%]

15: (mmhss→ hefrac) w3 [sup = 96.0% conf = 100%]

16: (heins→ hefrac,mmhss) w2 [sup = 96.0% conf = 100%]

17: (heins→ hefrac) w3 [sup = 96.0% conf = 100%]

BACCARS

1: (w1 → w2,w3) hefrac [sup = 96,0% conf = 100%]

2: (w1 → w2,w3) heins [sup = 96,0% conf = 98,0%]

3: (w1,w2 → w3) iadebt [sup = 90,0% conf = 100%]

4: (w1,w2 → w3) hecanb [sup = 90,0% conf = 97,8%]

5: (w1,w2 → w3) scqold [sup = 64,0% conf = 88,9%]

6: (w1,w2 → w3) scqoli [sup = 62,0% conf = 91,2%]

7: (w1 → w2) scqolf [sup = 74,0% conf = 86,0%]

8: (w1 → w2,w3) heins,hefrac [sup = 94,0% conf = 100%]

9: (w1 → w2) hefrac,mmhss [sup = 94,0% conf = 100%]

10: (w1 → w3) hefrac,iadebt [sup = 90,0% conf = 100%]

11: (w1→ w2) heins,hefrac,mmhss [sup= 92,0% conf= 100%]

12: (w1→ w3) heins,hefrac,iadebt [sup= 88,0% conf = 100%]

Discussion

It is possible to observe that the BCAAR type rules correspond to more descriptive rules that relate aspects present in long-lived profiles. For example, Rule 1 states that people who reported not having a hip fracture problem (hefrac) were not prevented from performing the walk test (mmhss). This occurred in 100% of the cases, during wave 2. This behavior is repeated in wave 3 (Rule 6), in 96% of cases that did not have hip fractures.

In relation to the BCAAR type rules, these rules are more predictive, which relate recurring situations in different waves. For example, Rule 1 states that long-lived individuals who did not have hip fractures did not present this problem along the 3 waves. The same interpretation occurs for the injection of insulin (Rule 2).

As another example, rule 5 establishes that long-lived individuals who do not have the feeling of abandonment during the first two waves, at the end of the third wave, they did not present this feeling, corresponding to 88.9% of the cases. That is, a feeling of abandonment was introduced in some individuals in the last wave, observed by the reduction in the confidence measure, which can affect the continuity of their longevity.

Conclusions

The results corroborate with existing studies in the literature, that the social, economic and family environment determine a complex structure directly connected to human aging.

In this work we bring a contribution to the analysis of aspects that favor longevity through triadic rules. We show that these rules allow to explore the temporal relationship of different aspects in longitudinal studies.

Finally, we emphasize that this approach can be extended to other clinical studies for which it is important to investigate the after-effects of a previous clinical procedure.

Acknowledgements

The data were made available through the UK Data Archive. ELSA was developed by a team of researchers based at the NatCen Social Research, University College London and the Institute for Fiscal Studies. The data were collected by NatCen Social Research. The funding is provided by the National Institute of Aging in the United States, and a consortium of UK government departments co-ordinated by the Ofice for National Statistics. The developers and funders of ELSA and the Archive do not bear any responsibility for the analyses or interpretations presented here.

This work was conducted during a scholarship supported by the National Council for Scientific and Technological Development of Brazil (CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnolgico) and CAPES, Brazilian Federal

Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil. This work was carried out at the Pontifical Caholic University of Minas Gerais, PUC-Minas.

References

[1] UNDESA. United Nations Department of Economic & Social Affairs. World population prospects: The 2017 revision, key findings and advance tables. Working Paper, No. ESA/P/WP. 241., 2017.

[2] L. Malloy-Diniz, D. Fuentes, and Cosenza, R. *Neuropsicologia do Envelhecimento: Uma Abordagem Multidimensional, 1st.Edition,*Artmed; Medicina e Saúde edition (January 1, 2013). In Portuguese Brazilian, ISBN-10 : 8582710143.

[3] C. E. Ribeiro, and L. E. Zárate, L. E. Data preparation for longitudinal data mining: a case study on human ageing. *Journal of Information and Data Management* 7 (2017), 116, 2017.

[4] B. Ganter, and R. Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer-Verlag New York, Inc., 1997. ISBN 3540627715.

[5] C. Carpineto, and G. Romano. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, 2004. ISBN 0470850558.

[6] F. Lehmann, and R. Wille. *A triadic approach to formal concept analysis. Conceptual structures: applications, implementation and theory*, Springer, 1995.

[7] K. Biedermann. How triadic diagrams represent conceptual structures. *Inter. Conf. on Conceptual Structur*es, ICCS 1997, p.304{317, 1997.

[8] B. Ganter, and S. Obiedkov. Implications in triadic formal contexts. *Inter. Conf. on Conceptual Structur*es, ICCS 2004, p.186{195, 2004.

[9] R. Missaoui; and L. Kwuida. Mining triadic association rules from ternary relations. *Inter. Conf. on Formal Concept*, Springer, p. 204-218, 2011.

[10] C. Ribeiro, L. E. Zárate. Classifying longevity profiles through longitudinal data mining. *Expert Systems with Applications*, v. 117, 09 2018.

[11] S. Minhas, A. Khanum, F. Riaz, A. Alvi, S.A. Khan. Early alzheimer's disease prediction in machine learning setup: Empirical analysis with missing value computation. In *Intern. Conf. on Intelligent Data Engineering and Automated Learning*. Springer, pp. 424–432, 2015.

[12] C.E. Ribeiro, L.H.S. Brito, C.N. Nobre, A.A. Freitas, and L.E. Zárate. A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7 (3), 2017.

[13] J. Grobelny, R. Michalski, Various approaches to a human preference analysis in a digital signage display design, Human Factors and Ergonomics in Manufacturing & Service Industries 21 (6) (2011) 529–542.

[14] B. G. Tabachnick, L. S. Fidell, J. B. Ullman, Using multivariate statistics, Vol. 6, Pearson Boston, MA, 2013.

[15] Rokia Missaoui and Kevin Emamirad (2017). Lattice Miner 2.0: A Formal Concept Analysis Tool (2017). Supplementary Proceedings of ICFCA'2017, Rennes, June 2017, p. 91-94.

[16] ELSA-UK. English Longitudinal Study of Ageing Available in: http://www.elsa-project.ac.uk

Address for correspondence

Marta D. M. Noronha - martadmnoronha@gmail.com
Luis E. Zárate – zarate@pucminas.br