Expert Systems With Applications 117 (2019) 75–89

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Classifying longevity profiles through longitudinal data mining

Caio Eduardo Ribeiro, Luis Enrique Zárate*

Department of Computer Science Applied Computational Intelligence Laboratory-LICAP Pontifical Catholic, University of Minas Gerais, 255 Walter Ianni st. São Gabriel, Belo Horizonte, MG 31980-110, Brazil

ARTICLE INFO

Article history:
Received 7 February 2018
Revised 14 September 2018
Accepted 15 September 2018
Available online 19 September 2018

2018 MSC:
00-01
99-00

Keywords:
Machine learning
Longitudinal data
Cluster analysis
Ageing studies

ABSTRACT

Populational studies of human ageing often generate longitudinal datasets with high dimensionality. In order to discover knowledge in such datasets, the traditional knowledge discovery in database task needs to be adapted. In this article, we present a full knowledge discovery process that was performed on a longitudinal dataset, mentioning the singularities of this process. We investigated the English Longitudinal Study of Ageing's (ELSA's) database, employing both semi-supervised and supervised learning techniques to determine and describe the profiles of individuals annotated with the class labels "short-lived" and "long-lived" who participated in the study. We report on the data preprocessing, the clustering task of finding the best sets of representatives of the profiles of each class, and the use of supervised learning to describe these profiles and perform a longitudinal classification on the dataset to investigate how consistently the unlabelled records would fit into the classes. The results show that several aspects are used to discriminate the individuals between the longevity profiles. Those aspects include economic, social and health-related attributes. The findings have pointed towards a need to further investigate the relationships between the different aspects, especially those related to physical health and wellbeing, and how they affect the lifespan of an individual. Furthermore, our methodology and the adopted procedures can be applied to any other data mining applications for longitudinal studies of ageing.

1. Introduction

Longitudinal data is a form of temporal data, in which the same samples are observed repeatedly at different points in time, called waves. Applying a Knowledge Discovery in Databases (KDD) process to a longitudinal dataset allows for a cause and effect investigation, due to the evolutionary aspect of the data. In a longitudinal dataset it is possible to analyse the sample's characteristics before and after an event occurs, and thus to infer the impact that event had on the samples. This makes these datasets highly recommended to investigate the impact of the passage of time (Diggle, Heagerty, Liang, & Zeger, 2002).

Frequently, populational studies observe individuals that share common characteristics, such as age, gender and/or social conditions, across the study's waves. A type of populational study that has been gaining attention from the scientific community and governmental agencies are human ageing populational studies. These studies investigate the evolution of aspects from several dimensions, such as social, economic, health and well-being, aiming to characterise individuals as they age, and discover knowledge about the ageing process.

It is estimated that the elderly population will surpass 21.5% of the global population by 2050 (United Nations Department of Economic & Social Affairs, 2017), impacting societal structures, with weighty social and economic implications (Lutz, Sanderson, & Scherbov, 2008). Both genetics and the environment play a role on the way an individual ages and, by understanding and influencing these roles, it is possible to increase the longevity, productivity and well-being of individuals, on a populational scale.

In a previous research study (Ribeiro, Brito, Nobre, Freitas, & Zárate, 2017), the researchers were unable to find articles that reported on the applications of unsupervised machine learning algorithms to perform longitudinal analysis on datasets of ageing studies. From the literature review, it was observed that the most frequent methods that were applied to perform longitudinal analyses on human ageing datasets were regression analysis, hypothesis tests, and correlation analysis. These classical statistical analysis methods are usually parametric, making strong assumptions about the data's distribution. Moreover, they often only detect linear correlations in the data. The study also determined that, among the most prominent longitudinal ageing studies, the English Longitudinal Study of Ageing (ELSA)[1] was the most complete one for

* Corresponding author.
E-mail address: zarate@pucminas.br (L.E. Zárate).

[1] English Longitudinal Study of Ageing website:http://www.elsa-project.ac.uk.

https://doi.org/10.1016/j.eswa.2018.09.035
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

a general-purpose research, because it investigates most of the aspects that are usually deemed to be important for ageing researches.

In this article, we report a full process of KDD process using longitudinal data. The process refers to an investigation of the ELSA database, aiming to describe the profiles that characterise two different classes: individuals who reached a longevity status during the study (long-lived class label), and those that died before reaching longevity (short-lived class label). This KDD process has been thoroughly described, including the preprocessing that was performed on the dataset, detailing a conceptual feature selection task and the transformations of the attribute values, considering the longitudinal aspect of the data.

Frequently, longitudinal studies have highly similar participants, as well as high dimensionality in their databases, which hinders the task of finding patterns. There are two main approaches for dealing with these poorly delineated regions: a) adapted algorithms with a greater sensitivity, or a sensitivity which adapts to each variable in the dataset (Adhikari et al., 2015); and b) combining traditional algorithms in an attempt to detect and delimit the representative regions in the dataset (Niemann, Hielscher, Spiliopoulou, Völzke, & Kühn, 2015). We have found that the ELSA database has extremely low variability among its records, which makes the task of identifying patterns for the long-lived and the short-lived individuals challenging. In order to address this issue of differentiating among similar records, we have opted for an approach that was based on a combination of two traditional clustering algorithms.

In order to achieve our goal of describing the profiles of short-lived and long-lived ELSA respondents, a three-step approach was designed. Firstly, we employed clustering algorithms, following the semi-supervised learning paradigm (Grira, Crucianu, & Boujemaa, 2004), to partition the records and find clusters with a majority of either long-lived or short-lived respondents. Secondly, the best clusters were validated using statistical analysis. Thirdly, we used these best clusters as a training set for a supervised learning algorithm, to train a classification model. Analyses of both the clusters and the classification results provided insights into the sought longevity-profiles.

Our approach is justified by the following: because of the complexity of human ageing and mortality, the study could not simply investigate every individual annotated with the long-lived class label as a member of the long-lived profile. The same applies to the short-lived class label. Therefore, it would not be adequate to apply a supervised learning process to an unprocessed dataset consisting of all long-lived and short-lived individuals in the ELSA database. We deemed it necessary to, first, cluster the long-lived and short-lived individuals to find a core group of representatives of each class, to make sure that the described profiles were accurate and relevant.

We believe that the profiles of long-lived and short-lived individuals are meaningful towards a better understanding of longevity and how it is influenced by environmental factors, such as economic and social characteristics. Our results show that several seemingly unrelated factors have appeared to have had an impact on whether an individual outlived their life expectancy or not. This corroborates with a well-known hypothesis that the environment is an intricate and complex structure, which is deeply connected to longevity (Cacioppo, Hughes, Waite, Hawkley, & Thisted, 2006; Kim, Sargent-Cox, French, Kendig, & Anstey, 2012). Therefore, this creates a demand for further studies on the relationships among ageing-related variables.

The main contributions in this article are as follows: a) to the best of the authors' knowledge, this is the first article reporting on a full, semi-supervised, KDD process applied to a longitudinal dataset with human ageing data.; b) we fully described the data preprocessing tasks that were performed in the ELSA database (which are all replicable for similar studies) to generate a longitudinal dataset; c) the results of our analysis have contributed towards profiling the long-lived and short-lived classes, by identifying the most relevant aspects that were highly apparent in differentiating them.

Another relevant contribution that can be highlighted is that selection of the “best examples” of a class, prior to applying the supervised machine learning to the dataset. When the records with known labels can not all be considered viable examples of a class, before any analyses are performed, employing semi-supervised learning to the dataset on the labelled instances can refine the sets of examples, as shown in our study. In human ageing studies and in Humanities studies in general, there are several applications that could benefit from this approach of refining a training set (Zhou & Belkin, 2014), such as the profiling of residents in a given location. There are several other longitudinal studies with databases that can be used in such applications, such as The Survey of Health, Ageing and Retirement in Europe (SHARE) and the Chinese Longitudinal Healthy Longevity Survey (CLHLS).

This article is organised as follows. Section 2 contains background information related to our approach. In Section 3, some related works are presented and we discuss how our study has differed from them. Section 4 contains a description of the methodology and its first two phases: data preprocessing and clustering. Section 5 presents the classification results and analyses of the longevity profiles, which refer to the final phase of the methodology. Section 6 finalises the article with our conclusions and, as unsupervised machine learning is rarely applied to the longitudinal analysis of ageing datasets, we recommended some interesting research topics that we believe would also be worthwhile of being investigated.

2. Background

2.1. Semi-supervised learning

As mentioned previously, in order to describe the profiles of the long-lived and short-lived ELSA participants, clustering techniques were employed, following the semi-supervised learning paradigm.

As defined by Grira et al. (2004), semi-supervised clustering algorithms can be of one of three types: a) algorithms that are based on learning through metrics, which aim to approximate the records by modifying the similarity measures; b) restriction-based algorithms, which apply restrictions that influence the adding of records to a cluster that is formed by the algorithm; and c) label-based algorithms, which consider that part of the records are labelled with the group to which they should belong.

In this study, we have particularly considered the label-based algorithms category, since the ELSA database has records of both the long-lived and short-lived classes. The study's goal was to find, among these labelled records, the best representatives of the long-lived and short-lived profiles. The class labels were used in order to evaluate cluster quality.

2.2. Longitudinal datasets

When creating a dataset from the ELSA database, each question was encoded as an attribute, and the attributes were divided into categories, such as economic, social, health-related and well-being attributes. The dataset originated from the ELSA database (Fig. 1)