A Mini-Survey of Supervised Machine Learning Approaches for Coping with Ageing-Related Longitudinal Datasets

Caio Ribeiro, Alex A. Freitas
University of Kent, Canterbury, United Kingdom
cer28@kent.ac.uk, A.A.Freitas@kent.ac.uk

Abstract

Longitudinal datasets have an added time index in their features, representing repeated measures of the same variables. This added temporal information increases the complexity of data representation, and carries meaningful information about the evolution of a feature's values throughout the time a study is conducted. This article introduces a new taxonomy of machine learning approaches to cope with longitudinal data, which are categorised as data transformation approaches and algorithm-adaptation approaches. Data transformation techniques transform the longitudinal data into a format that can be used by standard machine learning algorithms, possibly losing the time information in the data. Conversely, algorithm adaptation approaches change existing machine learning techniques to make them exploit the temporal information present in the input longitudinal data.

1 Introduction

Longitudinal datasets, where each data instance has variables repeatedly measured across multiple time points (called 'waves'), are increasingly important in many areas, particularly in biomedical science, due to the strong interest in understanding a disease's evolution across the lifetime of individuals. Longitudinal databases of ageing [Kaiser, 2013], [Ribeiro *et al.*, 2017] are particularly relevant for this, since old age is the greatest risk factor for a number of diseases.

However, although there are many supervised machine learning (ML) algorithms for classification and regression, few of them can directly cope with longitudinal datasets, and there is a general lack of survey papers discussing how supervised ML algorithms can cope with longitudinal data (some exceptions are discussed in Section 2).

In this context, this work presents a short survey of studies where supervised ML methods have been applied to longitudinal ageing-related datasets – where usually the class or target variable to be predicted represents the occurrence of some age-related disease.

This work is organised as follows. In Section 2 we briefly review related surveys (or reviews) on this topic, including mainly a simple taxonomy for categorising ML studies applied to longitudinal data previously proposed in [Jie *et al.*, 2017] – the only taxonomy of this kind so far in the literature, to the best of our knowledge. In Section 3, we discuss the limitations of that simple taxonomy and then we propose a new, more detailed taxonomy. This taxonomy has two dimensions, the first one involving the distinction between data transformation and algorithm adaptation. The former approach consists of transforming longitudinal data into standard, non-longitudinal data, so that standard supervised ML algorithms can be applied (reviewed in Section 4). The latter approach involves modifying an existing ML algorithm to make it directly cope with (untransformed) longitudinal data (reviewed in Section 5). The second dimension distinguishes between different types of data transformation. The discussions in Sections 4 and 5 are summarised in Section 6.

By describing how different studies fit into the different categories of approaches associated with the new proposed taxonomy, we can hopefully get more insight about important similarities and differences, as well as the pros and cons, of different supervised ML techniques for analysing longitudinal ageing-related datasets, which could lead to the design of better ML algorithms for this type of data.

2 Related Surveys (or Reviews)

Kaiser (2013) has reviewed a number of longitudinal databases of human ageing, focusing on the original purpose and the types of data included in the databases, but without any significant discussion on ML methods for analysing such data. Ribeiro *et al.* (2017) proposed a taxonomy focused on categorising different types of ageing-related variables, rather than categorising supervised ML methods for coping with longitudinal data. Also, they review mainly the use of classical statistical methods (e.g. logistic regression), rather than highly non-linear machine learning methods.

The most related work to this current work is [Jie *et al.*, 2017], which categorises supervised ML methods that use data from multiple time-points into four categories, based on the number of input and output time-points used by the ML method: (1) Single-time-point Input and Single-time-point Output (SISO), (2) Single-time-point Input and Multiple-time-points Output (SIMO), (3) Multiple-time-points Input and Single-time-point Output (MISO), and (4) Multiple-time-points Input and Multiple-time-points Output (MIMO). In the terminology used in this work, the inputs are features, the outputs are target variables, and time points are waves. Hence, a

SISO dataset has a single wave with features and target variables. A SIMO dataset also has features from a single wave, but the target variables span multiple waves. A MISO dataset has features in multiple waves but target variables in a single wave (typically, the last wave). A MIMO dataset has both features and target variables available in multiple waves.

3 A New Taxonomy of ML Approaches to Cope with Longitudinal Data

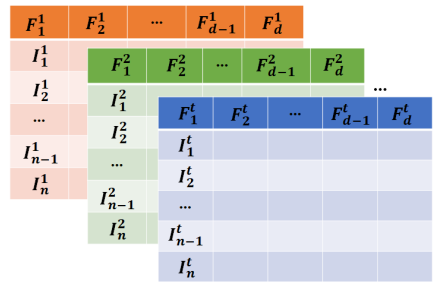
As discussed earlier, longitudinal datasets can be categorised based on the number of input and output time-points used by the ML method – i.e. the number of waves with features and target variable(s). However, this is a very high-level, coarse-grained categorisation, since it just refers to the cardinalities of the sets of features and class variables, without indicating how a ML algorithm would cope with the longitudinal nature of the underlying dataset. For successfully applying ML to such longitudinal datasets we need to choose a more specific approach to cope with the longitudinal nature of the data.

In this section we propose a new taxonomy of supervised ML approaches to cope with longitudinal data, which is more focused on how supervised ML algorithms can process longitudinal data. The proposed taxonomy has two dimensions. The first dimension involves a distinction between the data-transformation and the algorithm-adaptation approaches. In this approach, the longitudinal (temporal) data is first transformed into a format that can be directly used by standard supervised ML algorithms. Then, those algorithms are applied to the transformed data. The main advantages of this approach are its simplicity and generality, since it allows the use of many existing supervised ML algorithms. The disadvantage is that the data transformation usually loses some relevant temporal information about the data. Conversely, the algorithm-adaptation approach is more complex and is specific to each type of supervised ML algorithm, but it can potentially exploit better the full temporal information available in the original dataset.

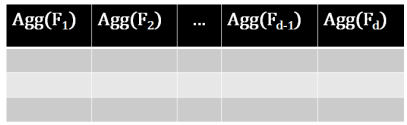
For the second dimension, we propose a categorisation of approaches for representing longitudinal features in a way that it is suitable for a standard (non-longitudinal) supervised ML algorithm. We focus on longitudinal features (inputs), as their representation is more commonly addressed in the literature than the representation of longitudinal class variables (outputs). Actually, the longitudinal ML studies reviewed in this work fall into the previously mentioned MISO or MIMO categories of [Jie *et al.*, 2017]'s taxonomy (both those categories contain features in multiple waves).

In order to describe feature representations, we use the following notation. Consider a longitudinal dataset with t consecutive waves $W_1, \ldots, W_t$, each of those comprised by d features and n instances (typically individuals, in longitudinal datasets of human ageing). As shown in Figure 1, we have identified four different approaches for coping with longitudinal (varying across time) features, in order to put the dataset into a format suitable for a supervised ML algorithm.

In the first approach (Figure 1(a)), denoted SepWav (Separate Waves), the ML algorithm receives as input one wave at a time. In the second approach, denoted AggrFunc, some aggregation function(s) – e.g., the mean or mode – is used to

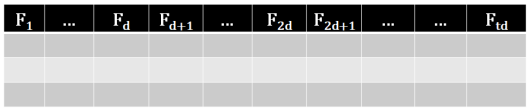


(a) SepWav: waves are inputted separately to the ML algorithm



(b) AggrFunc: aggregation functions summarise the values of a feature across all waves

(c) MerWav-Time(+): merged multiple waves, keeping the features' time indices



(d) MerWav-Time(-): merged multiple waves, discarding the features' time indices

Figure 1: Four feature representation approaches.

aggregate all values of a feature across multiple waves into a single value. This approach has the advantage of giving a lower dimensionality dataset, with summarised features, to the ML algorithm. However, this loses detailed information, and so the ML algorithm is unable to identify more detailed patterns of variation in a feature's values across time. By contrast, in the next two approaches, the ML algorithm receives as input the feature values from all waves, as follows.

The third approach, MerWav-Time(+), consists of merging multiple waves and keeping the features' time indices. This allows the use of a "time-aware" ML algorithm, i.e. a ML algorithm specifically designed or adapted to cope with longitudinal data where each feature value has a time index. Hence, using $f_{i,j}$ to denote the value of the i-th feature at the j-th wave (time index), the ML algorithm should consider the values of the same feature on different waves (say $f_{1,1}$ and $f_{1,2}$) as something conceptually different from the values of two different features on the same wave (say $f_{1,1}$ and $f_{2,1}$). How the algorithm should cope with these two types of feature values is up to the algorithm's designer; the point is that this representation provides the algorithm with detailed temporal information, allowing the design of truly longitudinal, time-aware ML algorithms. The fourth approach, MerWav-Time(–), consists of merging multiple waves and discarding the features' time indices, producing a timeless merged set of features. In this representation, the ML algorithm is forced to treat two values of the same feature in two different original waves in the same way it treats two values of different features in the same original wave. To avoid that the merged dataset have multiple features with the same feature index, since the time index (wave number) is removed in this rep-

resentation, one can rewrite the indices of all features after the wave merging process, using sequential incrementing, as shown in the column headings of the table in Figure 1(d).

The two dimensions of the proposed taxonomy are related as follows. In general, if the approach chosen in the first dimension is algorithm adaptation, it is natural to use the MerWav-Time(+) feature representation in the second dimension, since this is the only representation that gives the ML algorithm access to multiple waves with the features' time indices. By contrast, if the approach chosen in the first dimension is data transformation, one can use any of the three other feature representations in the second dimension.

4 A Review of Longitudinal ML Studies Using the Data Transformation Approach

[Zhang *et al.*, 2016] merged data from two consecutive waves into a single dataset, discarding the features' time indices (MerWav-Time(–) representation, Figure 1(d)). The authors employed a standard (non-longitudinal) decision tree algorithm to predict the Activities of Daily Living (ADL) status (healthy or disability) in a given wave of the dataset, using features from that wave and the previous wave. It is important to note that the authors used the value of the class label in the previous wave as a feature, which makes the prediction problem substantially easier. It seems intuitive that an individual that currently has ADL issues (difficulty to perform daily tasks) will likely still have them in the next wave.

[Mo *et al.*, 2013] also merged data from two waves using the MerWav-Time(–) representation. The class variable belonged to the first wave, so they used features from the second wave to predict the class in a past wave (a counter-intuitive prediction scenario). Their classification task was discriminating patients diagnosed with Alzheimer's disease from those diagnosed as cognitively normal, using an ensemble of 5 standard classifiers.

A different approach based on the MerWav-Time(–) representation was used by [Niemann *et al.*, 2015], who grouped instances considering the features observed in each wave, creating features related to clustering information in each wave (such as an instance's distance to a cluster's centroid, the cohesion and silhouette index of the instance) and how those changed in relation to previous waves. The features created from the clustering results were added to the dataset prior to feature selection, and the features from the 3 waves were merged and used for learning ignoring their time indices. Even though the constructed features consider some temporal information (when comparing clusters across waves), this information is considered before presenting the data to the ML algorithm, so the features' time indices are not available to the algorithm – i.e., the MerWav-Time(–) representation.

[Minhas *et al.*, 2015] used features from the first 6 waves of the Alzheimer's Disease Neuro-imaging Initiative study to predict the class label on wave 6, using standard SVM to predict a subject's conversion from the mild cognitive impairment class to the Alzheimer's disease class. In the first experiment, they used only the first wave's feature values for training (SepWav representation, Figure 1(a)). In the other experiments, they used two aggregation functions, namely the arithmetic mean and the median of each feature throughout the waves (AggrFunc representation, Figure 1(b)).

5 A Review of Longitudinal ML Studies Using the Algorithm-Adaptation Approach

[Adhikari *et al.*, 2015] created a new dataset from the Cardiovascular Health Study Cognition Study (CHS-CS) database. The created dataset had data from subjects of each age in the 65..98 range as waves, totalling 34 waves. For example, the wave for age 70 had data from all subjects in the CHS-CS study when they were 70, regardless of when that data was collected. Their model predicted the odds of either death or dementia (different models were trained for each prediction type) of a subject when they reach t + 10 years of age, where t is the subject's age at the last wave of the dataset. This work tackled the classification task by using a Lasso regression model [Tibshirani, 1996] that considered the features' time indices (MerWav-Time(+) representation, Figure 1(c)). The algorithm used two regularizers: a standard Lasso regularizer, which encourages overall sparsity in the coefficients of the active features (i.e. features with coefficient greater than 0 in the linear model) in each wave; and the fused Lasso regularizer [Tibshirani *et al.*, 2005], which encourages contiguity in the coefficients of the active features across waves.

Similarly, [Jie *et al.*, 2017] adapted the Lasso algorithm to predict Mini Mental State Examination and Alzheimer's Disease Assessment Scale-Cognitive Subscale scores using longitudinal magnetic resonance imaging data. They proposed a novel temporally-constrained group Lasso method, named tgLasso, which uses two weight smoothing techniques. The first is a fused smoothness term, where two weights for the same feature at adjacent waves have a small difference (like in the above fused Lasso). The second is a new output smoothness term, which requires that the model's outputs at two adjacent waves also have a small difference. In one experiment, four waves of data were used separately for regression (SepWav representation). In the other experiments, two or more consecutive waves were joined into a single dataset and the features' time indices were considered by the proposed tgLasso regulariser (MerWav-Time(+)). They tested predicting the scores in all waves, one at a time, using only the first wave's features, and gradually incremented the number of feature waves included in the dataset. The results showed that tgLasso significantly improved regression performance when compared with the standard and group Lasso methods.

In another algorithm-adaptation work, [Du *et al.*, 2015] extended a previous longitudinal classification SVM, LSVC [Chen and DuBois Bowman, 2011], by making it a longitudinal regression algorithm. LSVC extends SVM to longitudinal data by simultaneously estimating the traditional SVM separating hyperplane parameters with the proposed temporal trend parameters, taking into account dependencies within subjects; and the same principle was used to derive the longitudinal regression SVM. They created two types of datasets, the first, using the MerWav-Time(+) representation, i.e, merging the data from multiple waves and keeping the features' time indices, which were used to calculate temporal trend parameters. The second approach created a new dataset with

Table 1: Summary of the main characteristics of the reviewed studies

Study	Machine Learning (ML) task	Main type of ML algorithm	Jie *et al.*'s taxonomy	Proposed Taxonomy	
				Main approach	Feature representation
[Zhang *et al.*, 2016]	Classification	Decision tree	MISO	Data transformation	MerWav-Time(-)
[Mo *et al.*, 2013]	Classification	Ensemble of 5 classifier types	MISO	Data transformation	MerWav-Time(-)
[Niemann *et al.*, 2015]	Classification	RF, DT, NB, KNN	MISO	Data transformation	MerWav-Time(-)
[Minhas *et al.*, 2015]	Classification	SVM	MISO	Data transformation	SepWav, AggrFunc
[Adhikari *et al.*, 2015]	Classification	Fused Lasso	MIMO	Algorithm adaptation	MerWav-Time(+)
[Jie *et al.*, 2017]	Regression	Temporal Group Lasso	MIMO	Algorithm adaptation	SepWav, MerWav-Time(+)
[Du *et al.*, 2015]	Regression	Longitudinal SVR	MIMO	Algorithm adaptation	AggrFunc, MerWav-Time(+)
[Huang *et al.*, 2016]	Regression	Random Forest	MIMO	Algorithm adaptation	MerWav-Time(+)
[Radovic *et al.*, 2017]	Feature selection	Temporal mRMF	MISO	Algorithm adaptation	MerWav-Time(+)
[Pomsuwan and Freitas, 2017]	Feature selection	Modified CFS	MISO	Algorithm adaptation	MerWav-Time(+)

only the means of the feature values from all waves. The MerWav-Time(+) representation led to better results.

[Huang *et al.*, 2016]'s study aims to predict some longitudinal Alzheimer's Disease clinical scores. The model predicted the score for each individual in all the waves after the first wave, using features from the current and all past waves of the dataset as input (MerWav-Time(+)). They presented a Random Forest (RF) regression algorithm adapted for sparse regression. The proposed RF algorithm outperformed traditional RF and other popular regression methods: Lasso regression, Ridge regression, and SVM. The RF model with the best predictive accuracy started at the first wave and used its feature values to predict the score for the second wave, then incorporated the predicted feature score into the dataset, repeating this process until every wave's feature score prediction was incorporated. Hence, they used multiple instances of the MerWav-Time(+) representation, since in each run of the algorithm the feature representation consists of merging multiple waves and keeping the features' time indices.

A temporal variation of the minimum Redundancy-Maximum Relevance (mRMR) filter algorithm for feature selection was proposed by [Radovic *et al.*, 2017]. The temporal mRMR accepts as input a dataset with multiple waves and the features' time indices (MerWav-Time(+)). The time indices are used to calculate class-feature correlations, as well as similarities between the pairs of features' values, across waves. The study aimed to classify patients as symptomatic or asymptomatic using gene expression data.

[Pomsuwan and Freitas, 2017] merged several waves of data extracted from the ELSA (English Longitudinal Study of Ageing) database, maintaining the features' time indices (MerWav-Time(+) representation). The features were divided into groups; each group containing variations of the same base feature across time. They transformed the data so that each group would be small enough to be inputted into the exhaustive search version of the Correlation-based Feature Selection (CFS) method [Hall, 1999]. This work used biomedical features from three waves of the ELSA study, predicting whether individuals would develop an ageing-related disease in a later (future) wave. The strategy was tested for 10 different age-related diseases separately (10 binary classification problems), using decision-tree and Naive Bayes algorithms. Their proposed method showed an improvement over the standard CFS greedy forward search applied to all features (without dividing the features into groups) when tested with Nave Bayes, but did not significantly improve the results when tested with the decision tree algorithm.

6 Summary and Conclusions

Table 1 summarises the main characteristics of the reviewed studies, namely: the ML task addressed in the study, the main type of supervised ML algorithm(s) used, the categorisation of the study based on [Jie *et al.*, 2017]'s taxonomy (Section 2), and finally the categorisation of the study based on the two dimensions of the proposed taxonomy (Section 3): whether it used a data-transformation or algorithm-adaptation approach, and the feature representation approach used. Note that all the 6 algorithm-adaptation studies used the MerWav-Time(+) feature representation, which is consistent with the fact that this is the only representation preserving features from multiple waves with their time indices, providing the full longitudinal information for the ML algorithm. In some studies, experiments compared the MerWav-Time(+) representation against a non-longitudinal representation (SepWav or AggrFunc), and in general the former led to better results.

One limitation of this work is that we focused on longitudinal features only. Future work could propose a taxonomy for coping with longitudinal class variables.

References

[Adhikari *et al.*, 2015] Samrachana Adhikari, Fabrizio Lecci, James T Becker, Brian W Junker, Lewis H Kuller, Oscar L Lopez, and Ryan J Tibshirani. High-dimensional longitudinal classification with the multinomial fused lasso. *arXiv preprint arXiv:1501.07518*, 2015.

[Chen and DuBois Bowman, 2011] Shuo Chen and F DuBois Bowman. A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(6):604–611, 2011.

[Du *et al.*, 2015] Wei Du, Huey Cheung, Calvin A Johnson, Ilya Goldberg, Madhav Thambisetty, and Kevin Becker. A longitudinal support vector regression for prediction of als score. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1586–1590. IEEE, 2015.

[Hall, 1999] Mark Andrew Hall. *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato Hamilton, Hamilton, New Zealand, 1999.

[Huang *et al.*, 2016] Lei Huang, Yan Jin, Yaozong Gao, Kim-Han Thung, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of aging*, 46:180–191, 2016.

[Jie *et al.*, 2017] Biao Jie, Mingxia Liu, Jun Liu, Daoqiang Zhang, and Dinggang Shen. Temporally constrained group sparse learning for longitudinal data analysis in alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, 64(1):238–249, 2017.

[Kaiser, 2013] Angelika Kaiser. A review of longitudinal datasets on ageing. *Journal of Population Ageing*, 6(1-2):5–27, 2013.

[Minhas *et al.*, 2015] Sidra Minhas, Aasia Khanum, Farhan Riaz, Atif Alvi, Shoab A Khan, Alzheimer's Disease Neuroimaging Initiative, et al. Early alzheimer's disease prediction in machine learning setup: Empirical analysis with missing value computation. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 424–432. Springer, 2015.

[Mo *et al.*, 2013] Jue Mo, Sana Siddiqui, Stuart Maudsley, Huey Cheung, Bronwen Martin, and Calvin A Johnson. Classification of alzheimer diagnosis from adni plasma biomarker data. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 569. ACM, 2013.

[Niemann *et al.*, 2015] Uli Niemann, Tommy Hielscher, Myra Spiliopoulou, Henry Völzke, and Jens-Peter Kühn. Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In *Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on*, pages 121–126. IEEE, 2015.

[Pomsuwan and Freitas, 2017] Tossapol Pomsuwan and Alex A Freitas. Feature selection for the classification of longitudinal human ageing data. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 739–746. IEEE, 2017.

[Radovic *et al.*, 2017] Milos Radovic, Mohamed Ghalwash, Nenad Filipovic, and Zoran Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics*, 18(1):9, 2017.

[Ribeiro *et al.*, 2017] Caio Eduardo Ribeiro, Luis Henrique S Brito, Cristiane Neri Nobre, Alex A Freitas, and Luis Enrique 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.

[Tibshirani *et al.*, 2005] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Zhang *et al.*, 2016] Yuejin Zhang, Hengyue Jia, Aihua Li, Jianbing Liu, and Haifeng Li. Study on prediction of activities of daily living of the aged people based on longitudinal data. *Procedia Computer Science*, 91:470–477, 2016.