Nested Trees for Longitudinal Classification

Sergey Ovchinnik
University of Kent
Canterbury, UK
S.Ovchinnik@kent.ac.uk

Fernando Otero
University of Kent
Canterbury, UK
F.E.B.Otero@kent.ac.uk

Alex A. Freitas
University of Kent
Canterbury, UK
A.A.Freitas@kent.ac.uk

ABSTRACT

Longitudinal datasets contain repeated measurements of the same variables at different points in time. Longitudinal data mining algorithms aim to utilize such datasets to extract interesting knowledge and produce useful models. Many existing longitudinal classification methods either dismiss the longitudinal aspect of the data during model construction or produce complex models that are scarcely interpretable. We propose a new longitudinal classification algorithm based on decision trees, named Nested Trees. It utilizes a unique longitudinal model construction method that is fully aware of the longitudinal aspect of the predictive attributes (variables) and constructs tree nodes that make decisions based on a longitudinal attribute as a whole, considering measurements of that attribute across multiple time points. The algorithm was evaluated using 10 classification tasks based on the English Longitudinal Study of Ageing (ELSA) data.

CCS CONCEPTS

• **Computing methodologies → Classification and regression trees**; • **Information systems → Data mining**.

KEYWORDS

Classification, Longitudinal Data, Decision Trees

ACM Reference Format:
Sergey Ovchinnik, Fernando Otero, and Alex A. Freitas. 2022. Nested Trees for Longitudinal Classification . In *Proceedings of ACM SAC Conference (SAC'22)*. ACM, New York, NY, USA, 4 pages. https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

Longitudinal datasets contain repeated measurements of the same attributes taken at different time points. The repeatedly-measured attributes of longitudinal datasets are referred to as longitudinal attributes. An example of a longitudinal attribute is an attribute consisting of 4 measurements of a patient's blood cholesterol levels, taken once every 4 years.

Longitudinal data mining aims to utilize such datasets in order to extract interpretable and potentially useful knowledge or patterns hidden in the data. The main advantage of such datasets is their temporal nature, which can be utilized to make predictions based not only on a set of attribute values, but also on the trends that occur within them over time. Most importantly, the longitudinal nature of these datasets can be used to construct models that use previously recorded data to predict future events.

Note that longitudinal data should not be confused with time series data [1], even though both have a temporal nature. In the context of classification (supervised learning), time series data typically contain a single real-valued variable repeatedly measured across a large number of timepoints; whilst our target longitudinal datasets consist of a large number of both numerical and nominal variables repeatedly measured across a small number of timepoints.

Over the past few years, several real-world longitudinal data mining studies have been conducted [17, 15, 8, 2, 7]. Such studies can be split into two approaches: data transformation (pre-processing) and algorithm adaptation (specialized algorithms).

Several recent works on longitudinal classification [9, 19, 10] use a pre-processing step known as *flattening*, where a longitudinal dataset is 'flattened' into a non-longitudinal one by considering each repeated time-specific measure of an attribute as a different attribute. This removes the longitudinal aspect from the data and allows non-longitudinal classification algorithms to be used.

In this work, we focus on the less explored approach of algorithm adaptation, and we propose a new type of decision tree algorithm for longitudinal classification, since many decision tree models have the advantage of interpretability [13, 5]. Decision trees are in general transparent and their decision making can be described by a simple diagram that a user can inspect manually – as long as the tree is not too large – to understand the exact reasoning behind every prediction made by the model.

The proposed longitudinal decision tree algorithm preserves the longitudinal nature of the dataset by constructing a Nested Trees model, where each node of the decision tree represents an embedded decision tree that makes decisions based solely on the values of one longitudinal attribute (i.e., its values across multiple time points). The final model is composed by an outer decision tree made up of inner decision trees in each node, hence the name *Nested Trees*. This algorithm produces longitudinally-aware prediction models, analysing each longitudinal attribute as a whole rather than treating its temporal values independently. It also provides longitudinal classification models more interpretable than the models produced by conventional decision tree algorithms.

The remainder of the paper is structured as follows. Section 2 describes the proposed longitudinal classification algorithm. Section 3 describes the dataset and the experimental methodology used to evaluate the proposed algorithm. Section 4 presents the results of the computational experiments and their discussion. Section 5 presents the conclusions.

SAC'22, April 25 –April 29, 2022, Brno, Czech Republic

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/xx.xxx/xxx_x

2 THE PROPOSED LONGITUDINAL CLASSIFICATION ALGORITHM

The proposed Nested Trees algorithm is a classification algorithm designed to be used with longitudinal datasets. The algorithm constructs a classification model similar in structure to models learned by conventional decision tree algorithms. The difference is that, while conventional decision tree models have nodes that can be expressed as simple attribute-value tests, the Nested Trees algorithm constructs a model that uses inner (typically smaller) decision trees as nodes of an outer (often larger) tree, thus constructing a decision tree made of decision trees. Hence, the construction procedure uses a two-layer structure. On the outer layer, it constructs a decision tree where each node uses a single longitudinal attribute to make the split. On the inner layer – inside each node of the outer tree – it constructs an embedded decision tree that uses the different temporal values of a single longitudinal attribute, i.e., the attribute's values measured at different time points. Hence, it takes full advantage of the longitudinal information present on the data. An illustration of a Nested Trees model is shown in Figure 1.

The inner embedded decision trees construct small decision trees that use the longitudinal values of a single longitudinal attribute. These inner trees are constructed by a custom decision tree algorithm implementation which is a hybrid of CART[4] and C4.5[13] algorithms. It constructs a tree made of binary splits, using a greedy approach based on the Gini impurity metric to select the attribute to split the data in the current node; a minimum node size requirement and a maximum depth limit are used as both the stopping criteria and the pre-pruning methods to mitigate overfitting. Missing values are handled in a way similar to the C4.5 algorithm: when an instance has a missing value for an attribute used to split the data in the current node, each branch coming out from that node gets a fraction of the instance (represented by an instance weight in [0..1]) equal to the proportion of instances in the current node having the attribute value associated with that branch.

The outer layer construction process uses the same algorithm with a different attribute selection mechanism. While traditional decision tree algorithms construct each node by selecting the single best attribute to split the current data into subsets, the outer layer selects a longitudinal attribute as a whole, which comprises a group of time-specific attributes, and searches for an optimal decision tree that would split the current data using only the values of the attributes from that specific group. In other words, the outer layer treats each longitudinal attribute as a unit, instead of considering each repeated time-specific measure of an attribute as a different attribute. Each node of the outer tree contains an embedded tree produced by the inner tree construction algorithm using only the temporal values of the selected longitudinal attribute. Each leaf of a node's inner decision tree is used to create a branch and the outer layer construction process is repeated on each generated subset.

The main advantage of the Nested Tree algorithm is longitudinal awareness. There is currently no work in the literature where a decision tree used in longitudinal classification had any inherent longitudinal awareness. All previous approaches in this area used a combination of the flattening approach and a simple decision tree algorithm to create longitudinal predictive models. Such approaches result in the model treating the different values of the same longitudinal attribute as independent attributes, ignoring the longitudinal aspect of the data. In contrast, the proposed algorithm analyses each longitudinal attribute as a whole and does not separate its longitudinal values (different time points) from each other as is the case in the flattening approach. The longitudinal aspect of the data is preserved and the model consists of an outer decision tree, where each node represents a split on a whole longitudinal attribute instead of just one of its temporal values.

This has several advantages regarding model usability. The first advantage is the potential improvement in predictive accuracy as the algorithm is able to benefit from the longitudinal information.

The second advantage is the model interpretability and the apparent attribute importance. The trees produced using a flattening approach are not restricted in how they use the attributes in model construction, so they can produce models where different time-specific attributes representing different values of a single longitudinal attribute (e.g., values of the longitudinal cholesterol attribute measured across different time points) are scattered around different parts of the model. This makes it difficult to estimate attribute importance for longitudinal attributes, since a longitudinal attribute's impact on the prediction is not always clear. In the proposed approach, all temporal values of a longitudinal attribute are grouped together (in an inner decision tree) in a node of the outer tree,

thus making it easier to analyze the importance of a longitudinal attribute as a whole. Improved model acceptance by users is another advantage. A longitudinally-aware decision model that preserves the longitudinal nature of the data during model construction can also make the model more likely to be accepted by domain experts than a longitudinally unaware model.

In addition, the algorithm implementation uses instance weights that represent how much each instance affects the prediction. Hence, in order to mitigate class imbalance issues (where one class is much less frequent than another [6]), a class balancing pre-processing step is added to adjust the instance weights to make the total sum of instance weights equal for each class. This adjustment is made as a pre-processing step for the training set before model construction.

In the remainder of this section we discuss the most related work. The proposed Nested Trees algorithm bears some similarity to the Tree of Predictors algorithm (ToP) [18]. ToP also constructs a prediction model that has an outer structure resembling a decision tree with nodes representing embedded predictors. The main difference is that, while the Nested Trees algorithm always uses inner decision trees as nodes of the outer tree, the ToP algorithm has a larger variety of predictors (some of them not interpretable) that can be used as inner nodes. Also, ToP was not designed for longitudinal classification, it does not recognize longitudinal attributes.

The RE-EM trees algorithm [16] is a decision tree algorithm for clustered and longitudinal data. Similarly to Nested Trees, it has an algorithm-level adaptation for dealing with attributes that can be grouped together by some common property, e.g. different temporal measurements of the same longitudinal attribute. The RE-EM trees algorithm provides some additional functionality that can be used for constructing regression models using longitudinal datasets: clustering the attributes into groups and using mixed effects. However, it does not use a longitudinally-aware model construction process and does not produce a longitudinally-interpretable model.

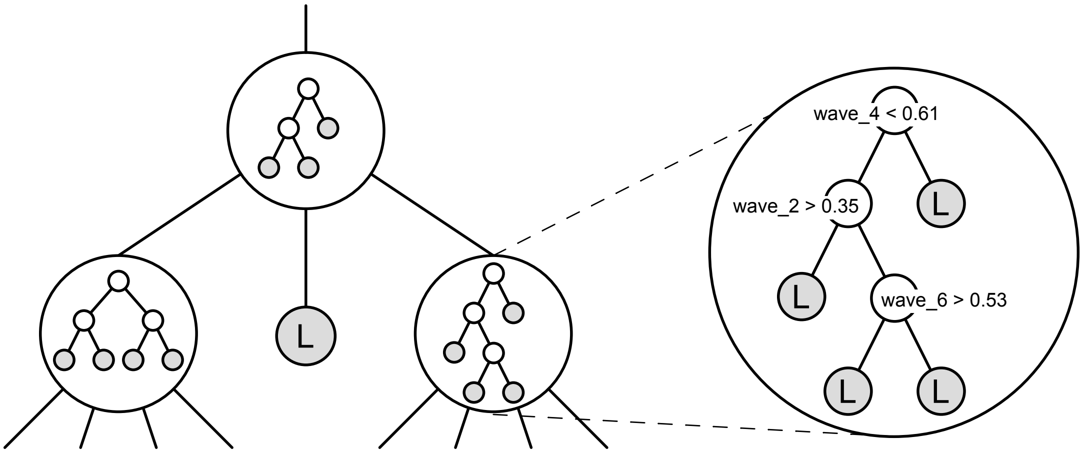

Figure 1: Illustration of a Nested Trees model. Internal nodes of the outer tree use values of a single longitudinal attribute to construct an inner decision tree, while leaf nodes—labelled as 'L'—correspond to class predictions. Each node of an inner tree makes a split based on the value of the longitudinal attribute at a specific time point (wave); leaf nodes of the inner trees correspond to the branches of the outer node the inner tree is nested in.

In [11], the XGBoost algorithm was used to learn boosted decision tree models from the same longitudinal datasets used in this current paper. That study focused on improving model acceptability by using monotonicity constraints to produce monotonic classification models, instead of improving the longitudinal awareness of the models. Hence, that study did not produce fully longitudinally-aware models and only used a set of derived attributes to represent longitudinal information.

Another work [14] proposed a random forest method for longitudinal classification, which selects features in tree nodes based on both their predictive power and their time indices (favouring more recent features). However, their random forest models are not directly interpretable.

3 EXPERIMENTAL METHODOLOGY

The dataset used in this study comes from the "Nurse Visit" section of the English Longitudinal Study of Ageing (ELSA) [3]. The predictor attributes represent patient health measurements such as blood test results and physical exercise tests. Ten class attributes are derived from the ELSA dataset, each expressed as a binary class variable representing presence or absence of a certain disease in the final wave (time point), i.e., wave 8. The 10 classification problems – one for each class attribute – contain records of the same 7097 individuals participating in the ELSA study.[1]

Each record contains 2 non-longitudinal attributes (age and sex) and 40 longitudinal attributes. Since the "Nurse Visit" data is available only for waves (time points) 2, 4, 6 and 8 of the ELSA study, each longitudinal attribute is represented by up to 4 separate attributes (one for each of those waves) and a class label. In total, the dataset contains 130 predictive attributes, counting all multiple values of each longitudinal attribute across the four waves. This dataset was used to create 10 different classification problems – all problems using the same set of 130 attributes, but each problem using a different disease as the class variable to be predicted.

A full description of attributes used and their meaning can be found in a related study that previously used the same data preparation techniques in the context of automatic feature selection [12]. The same dataset has also been previously used to evaluate other longitudinal data mining approaches [11].

A well-known 10-fold cross validation was used to evaluate the performance of the constructed models. Additionally, in each experiment, the cross-validation was repeated 30 times (varying the random seed) and the results were averaged over all runs. The cross-validation was applied to data instances only and not to the time points of longitudinal attributes – both training and test subsets used the full set of longitudinal time points.

In addition, the algorithm implementation uses instance weights that represent how much each instance affects the prediction. Hence, in order to mitigate class imbalance issues (where one class is much less frequent than another [6]), a class balancing pre-processing step is added to adjust the instance weights to make the total sum of instance weights equal for each class. This adjustment is made as a pre-processing step for the training set before model construction.

The experiments compared the proposed Nested Tree algorithm against a conventional Decision Tree algorithm (the same as the one used for constructing the inner trees of the Nested Trees models). These algorithms were evaluated in terms of two predictive accuracy measures: the average F-Measure values over the two class labels, and the average Area Under the ROC curve (AUROC).

Based on preliminary experiments we use the following hyperparameter settings for all of our experiments: Maximum Outer Tree Depth: 10; Maximum Inner Tree Depth: 5; Minimum Tree Node Size: 2 (for both inner and outer tree nodes).

[1] The codebase for this project and the instructions for accessing the dataset can be found at: http://github.com/NestedTrees/NestedTrees

Table 1: Comparison of average predictive accuracy measures of the two algorithms. Higher values are highlighted in boldface.

	F-Measure		AUROC	
Dataset	Decision Tree	Nested Tree	Decision Tree	Nested Tree
Angina	0.455	**0.515**	**0.537**	0.517
Arthritis	**0.560**	0.548	**0.572**	0.548
Cataract	**0.620**	0.575	**0.657**	0.577
Dementia	0.420	**0.532**	**0.687**	0.536
Diabetes	0.365	**0.584**	0.584	**0.595**
HBP	0.555	**0.602**	0.577	**0.604**
Heart attack	0.382	**0.513**	**0.565**	0.537
Osteoporosis	0.403	**0.541**	**0.573**	0.570
Parkinson's	0.345	**0.500**	**0.587**	0.508
Stroke	0.280	**0.527**	**0.576**	0.550
# wins	2	8	8	2

4 EXPERIMENT RESULTS AND ANALYSIS

Table 1 reports the average F-Measure and AUROC values for the two algorithms. The average F-measure was computed by considering each class in turn as the positive class and macro-averaging the results, i.e., considering both classes as equally important.

The Nested Tree algorithm achieved higher F-measure values in 8 out of 10 classification problems; but the conventional decision tree achieved higher AUROC values in 8 out of 10 problems.

These AUROC results are likely the result of the different sizes of the models generated by the two algorithms. The Nested Trees algorithm tends to create a large number of splits, resulting in a very large tree, where each of the leaf nodes uses a very small number of training instances (usually 2-4 instances). It can therefore only have a very small number of different class probability values, making the ROC curve less defined, having only a few points between [0,0] and [1,1] and thus having a smaller area under the curve than a smaller model generated by the conventional decision tree algorithm.

5 CONCLUSIONS

This work proposed a new longitudinally-aware Nested Trees algorithm that constructed a decision tree structure made of nodes that contained inner decision trees. This algorithm did not use the flattening pre-processing step and constructed models that properly took into account the longitudinal nature of the dataset.

The proposed algorithm outperformed a conventional longitudinally-unaware decision tree algorithm in terms of average F-Measure, but the latter outperformed the former in terms of the AUROC measure.

REFERENCES

[1] Anthony Bagnall et al. "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances". In: 31.3 (Nov. 2016), pp. 606–660.

[2] Marıa T Ballestar et al. "Impact of robotics on manufacturing: A longitudinal machine learning perspective". In: *Technological Forecasting and Social Change* 162 (2021).

[3] M. Blake et al. *English Longitudinal Study of Ageing: Waves 0-8, 1998-2017.* 2018.

[4] L. Breiman et al. *Classification And Regression Trees.* Routledge, 1984.

[5] Alex A. Freitas. "Comprehensible classification models". In: *ACM SIGKDD Explorations Newsletter* 15(1) (2014), pp. 1–10.

[6] Qiong Gu et al. "Data Mining on Imbalanced Data Sets". In: *2008 International Conference on Advanced Computer Theory and Engineering.* IEEE, 2008, pp. 1020–1024.

[7] Ann F. Haynos et al. "Machine learning enhances prediction of illness course: a longitudinal study in eating disorders". In: *Psychological Medicine* (2020), pp. 1–11.

[8] Samantha Joel, Paul W. Eastwick, and et al. Colleen J. Allison. "Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies". In: *Proc. of NAS* 117(32) (2020), pp. 19061–19071.

[9] Jue Mo et al. "Classification of Alzheimer Diagnosis from ADNI Plasma Biomarker Data". In: *Proc. of the International Conf. on Bioinformatics, Computational Biology and Biomedical Informatics.* ACM, 2013, pp. 569–574.

[10] Uli Niemann et al. "Can We Classify the Participants of a Longitudinal Epidemiological Study from Their Previous Evolution?" In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems.* 2015, pp. 121–126.

[11] Sergey Ovchinnik, Fernando E. B. Otero, and Alex A. Freitas. "Monotonicity Detection and Enforcement in Longitudinal Classification". In: *LNCS, Artificial Intelligence XXXVI.* Springer International Publishing, 2019, pp. 63–77.

[12] Tossapol Pomsuwan and Alex A. Freitas. "Feature Selection for the Classification of Longitudinal Human Ageing Data". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW).* 2017, pp. 739–746.

[13] J.R. Quinlan. *C4.5 : programs for machine learning.* San Mateo, Calif. : Morgan Kaufmann Publishers, 1993.

[14] Caio Ribeiro and Alex Freitas. "A New Random Forest Method for Longitudinal Data Classification Using a Lexicographic Bi-Objective Approach". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI).* 2020, pp. 806–813.

[15] Caio Ribeiro and Alex A. Freitas. "A Mini-Survey of Supervised Machine Learning Approaches for Coping with Ageing-Related Longitudinal Datasets". In: *3rd Workshop on AI for Aging, Rehabilitation and Independent Assisted Living (ARIAL), IJCAI-2019.* 2019.

[16] Rebecca J. Sela and Jeffrey S. Simonoff. "RE-EM trees: a data mining approach for longitudinal and clustered data". In: By Machine Learning, 86(2), 2012 (July 2011), pp. 169–207.

[17] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin. "Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning". In: *Journal of Child Psychology and Psychiatry* 59(12) (2018), pp. 1261–1270.

[18] Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. "ToPs: Ensemble Learning With Trees of Predictors". In: *IEEE Transactions on Signal Processing* 66.8 (2018), pp. 2141–2152.

[19] Yuejin Zhang et al. "Study on Prediction of Activities of Daily Living of the Aged People Based on Longitudinal Data". In: *Procedia Computer Science* 9 (2016), pp. 470–477.