Chapter 16
Efficient End-to-End Asynchronous Time-Series Modeling With Deep Learning to Predict Customer Attrition

Victor Potapenko
Florida International University, USA

Malek Adjouadi
Florida International University, USA

Naphtali Rishe
Florida International University, USA

ABSTRACT

Modeling time-series data with asynchronous, multi-cardinal, and uneven patterns presents several unique challenges that may impede convergence of supervised machine learning algorithms, or significantly increase resource requirements, thus rendering modeling efforts infeasible in resource-constrained environments. The authors propose two approaches to multi-class classification of asynchronous time-series data. In the first approach, they create a baseline by reducing the time-series data using a statistical approach and training a model based on gradient boosted trees. In the second approach, they implement a fully convolutional network (FCN) and train it on asynchronous data without any special feature engineering. Evaluation of results shows that FCN performs as well as the gradient boosting based on mean F1-score without computationally complex time-series feature engineering. This work has been applied in the prediction of customer attrition at a large retail automotive finance company.

DOI: 10.4018/978-1-7998-7156-9.ch016

INTRODUCTION

Time-series (TS) data can be classified as synchronous vs. asynchronous, co-cardinal vs. multi-cardinal, and even vs. uneven (Wu et al., 2018). Synchronous TS data are aligned on the time dimension, while asynchronous events are not. Co-cardinal event sequences are aligned on the first and last elements. TS are classified as multi-cardinal if one of the sequences is longer than another. Multi-cardinal sequences are synchronous if the remaining elements are aligned. TS are defined to be even when data points are distributed evenly over time. Unless otherwise specified, we use asynchronous as a general term to describe asynchronous, multi-cardinal, and uneven sequences.

Our instantiation of this problem is a supervised multi-class classification with stationary and TS data. As a case study, we use a dataset comprising more than one million car loan histories. Our TS are asynchronous, multi-cardinal, and uneven. Stationary data is comprised of context data, such as vehicle and customer profiles. The target of the classification task in this case study is the customer attrition classified by type of contract termination within a 6-month horizon. The proposed approach includes database TS data extraction, imbalanced target sampling, time-series dimensionality reduction for XGBoost models, originally proposed by Chen and Guestrin (2016), that do not support multi-dimensional data, and supervised classification using a Fully Convolutional Network (FCN). (XGBoost is an open-source software library that provides a gradient boosting framework.)

TS extraction is performed with a technique called "streaming updates." The streaming updates method is used to extract value updates as TS data at variable time intervals defined by a frequency of database of field updates. This approach is discussed in Section 3a on Data Extraction. At any given time step, our extracted data is significantly imbalanced towards active accounts that are not terminated. In Section 3b, we outline a sampling approach that mitigates class imbalance in training set and prevents data leakage into the evaluation set by using different sampling techniques designed to simulate real-world conditions.

The input size and frequency of TS varies widely among features extracted with the streaming method. Some features frequently occur at regular time intervals throughout account lifetimes; others only occur a few times at highly irregular intervals. As a result, the products of join operations cause input space to increase dramatically, which leads to high memory and computational requirements for their manipulation and creates the necessity to use imputation techniques to eliminate sparsity. The scope of the present work is limited to the comparison of the non-deep learning approach that uses TS statistics to reduce time dimension, versus the deep learning approach with FCN trained on zero-imputed TS data. In Section 3c, we describe the XGBoost and FCN models. Section 4 outlines evaluation results followed by a discussion in Section 5.

BACKGROUND

Time-series sequences have been modeled with Hidden Markov Models (HMMs) (Rabiner, 1989; Ephraim & Merhav, 2002) and Bayesian Networks (BNs) (Heckerman et al., 1995; Nielsen et al., 2009). However, HMMs and BNs are not designed for asynchronous sequences because they require specification of a constant time interval between consecutive events (Wu et al., 2018). Asynchronous data needs to be reshaped and synchronized to fit HMMs and BNs. TS can be reshaped to synchronicity at the data preprocessing step. Reshaping and synchronization methods often obfuscate original data or create artificial data points that are not initially present in the dataset. This results in information loss and requires

Related Content