Programa de Verão do LNCC

Jornada de Ciência de Dados

Apresentação do Docente

CEFET/RJ

Eduardo Ogasawara
eogasawara@ieee.org
https://eic.cefet-rj.br/~eogasawara

Biografia

- Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ) em 2011
- Docente do CEFET/RJ
  - Departamento de Ciência da Computação
  - Coordenação do Curso Técnico de Informática
- Docente permanente
  - Programa de Pós-graduação em Ciência da Computação (PPCIC)
  - Programa de Pós-graduação em Eng. de Produção e Sistemas (PPPRO)
- Membro da SBC, ACM, IEEE e INNS

https://eic.cefet-rj.br/~eogasawara
http://lattes.cnpq.br/0528303491410251
https://www.researchgate.net/profile/Eduardo_Ogasawara
https://www.linkedin.com/in/eogasawara

Produção Científica: Centrada em Mineração de Dados e Workflows

- 125 artigos publicados*
  - 25 de periódicos

ELSEVIER

Ecological Informatics
Volume 36, November 2016, Pages 94-105

Evaluating temporal aggregation for predicting the sea surface temperature of the Atlantic Ocean

ELSEVIER

Knowledge-Based Systems
Volume 164, 15 January 2019, Pages 274-291

Nonstationary time series transformation methods: An experimental review

Rebecca Salles [a], Kele Belloze [a], Fabio Porto [b], Pedro H. Gonzalez [a], Eduardo Ogasawara [a]

Show more

https://doi.org/10.1016/j.knosys.2018.10.041 Get rights and content

ELSEVIER

Future Generation Computer Systems
Volume 68, March 2017, Pages 111-127

Deriving scientific workflows from algebraic experiment lines: A practical approach

ELSEVIER

Transportation Research Part E: Logistics and Transportation Review
Volume 95, November 2016, Pages 282-298

An analysis of Brazilian flight delays based on frequent patterns

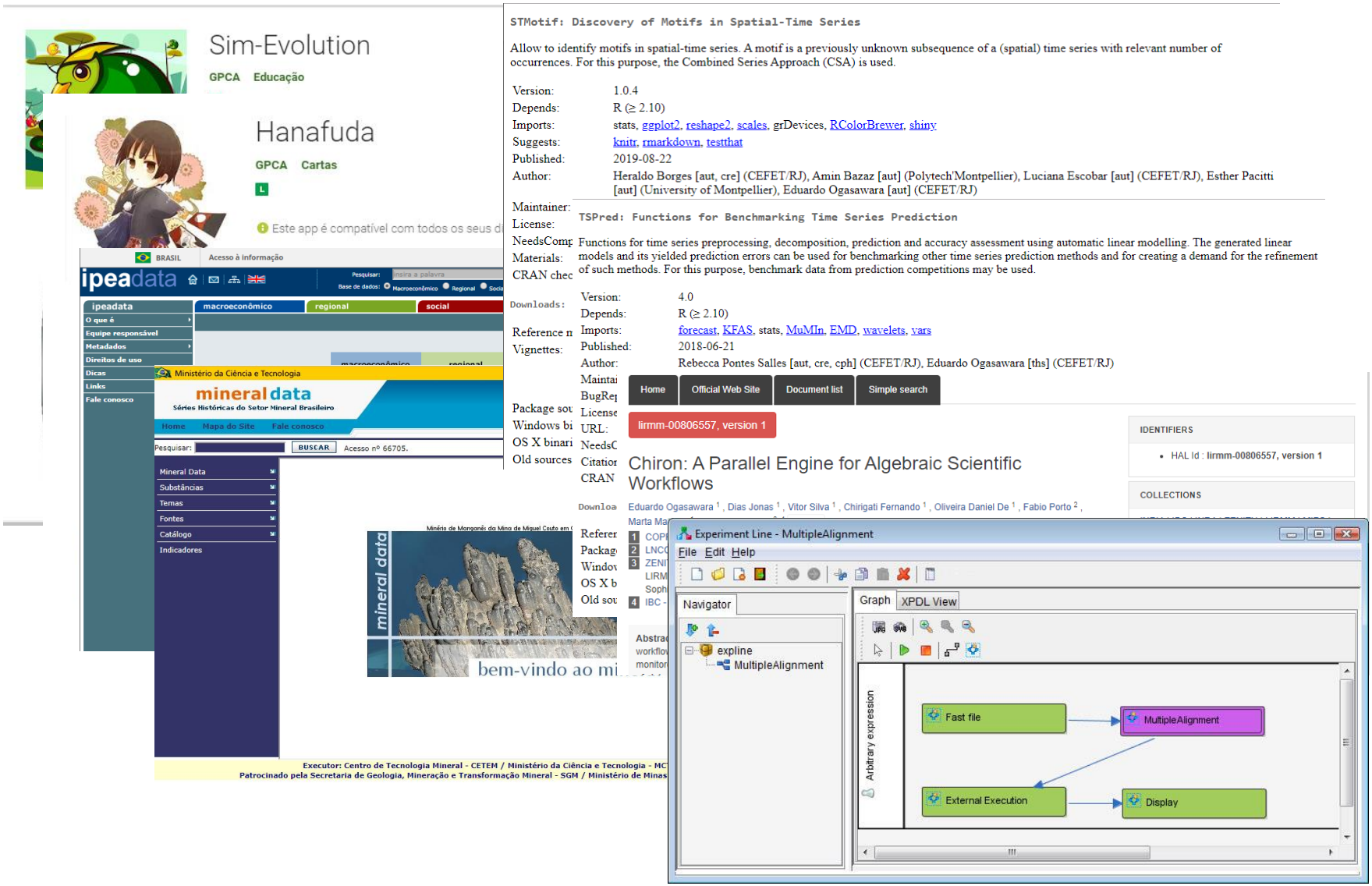
Alice Sternberg [a], Diego Carvalho [a], Leonardo Murta [c], Jorge Soares [a, b], Eduardo Ogasawara [a]

Show more

https://doi.org/10.1016/j.tre.2016.09.013 Get rights and content

* Revisor de periódicos e conferências internacionais

Produção técnica:
Diversos artefatos computacionais



Temas de Pesquisa: Predictive analytics

- Classificação
- Regressão
- Análise de séries temporais e espaço-temporais
- Não-estacionariedade
  - Descoberta e tratamento para *concept-drift*

On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays

Leonardo Moreira, Christofer Dantas, Leonardo Oliveira — CEFET/RJ — {leonardo.moreira,christofer.dantas,leonardo.oliveira}@eic.cefet-rj.br
Jorge Soares — CEFET/RJ — jorge@eic.cefet-rj.br
Eduardo Ogasawara — CEFET/RJ — eogasawara@ieee.org

***Abstract*—Flight delays cause various inconveniences for airlines, airports, and passengers. According to data provided by the Brazilian National Civil Aviation Agency (ANAC), between 2009 and 2015, about 22% of domestic flights made in Brazil were delayed by more than 15 minutes. The prediction of these delays is fundamental to mitigate their occurrence and optimize the decision-making process of an air transport system. Particularly, airlines, airports, and users may be more interested in when delays are likely to occur than the accurate prediction of the absence of delays. This paper focuses on the unbalanced distribution of the classes of delay (presence and absence) by performing an experimental evaluation of several preprocessing methods for the development of machine-learning flight delay classification models. Those models were built from a dataset that integrates national flight operations with meteorological conditions of airports. Our results indicate the models that applied the balancing techniques performed much better in predicting the occurrence of delays, getting about 60% of hits.**

I. Introduction

Delay is one of the key performance indicators of any transportation system. A flight delay shall be represented by the difference between the programmed time and the actual time of departure or arrival of a flight. In the context of commercial aviation, these delays can occur for a variety of reasons, including flight process failures, weather conditions, mechanical problems, ground delays, air traffic control and capacity constraints.

In the commercial aviation scenario, delays have a high financial impact on airlines, such as fines, additional operating costs, and declining customer loyalty. Also, given the uncertainty of its occurrence, many passengers are forced to reschedule their travels to arrive at the destination on time, which often leads to increased travel costs [1].

Specialized literature shows that large volume of data has been collected in databases of public and private institutions to study and to understand the operations of the air transport system. Analysis of this data is crucial for gaining the knowledge needed to detect and predict delays. Many recent studies have been done in analyzing flight data using machine learning methods [2], [3], [4]. Such initiatives are done thanks to the large volume of flight that has been collected in these years, in what is currently known as the Data Science era [5], [6], [7], [8].

Thus, methods of predicting flight delays are fundamental to mitigate their occurrence, and, as a consequence, reduce the financial losses. Therefore, classification models for predicting when delays may occur are needed, given the complexity of reasons and conditions that generate delays [9]. In fact, any improvement in this theme may be beneficial for airlines, airports, and passengers. Notably, the sensitivity of predicting when delays are likely to occur is more relevant than trying to target the accuracy of predicting the presence or absence of delays.

In this context, the objective of this work is to perform an experimental evaluation of data preprocessing methods for machine learning classification models, considering the factors involved and collected by the databases focusing on the sensitivity of machine learning classifiers. For that, this work builds a dataset that integrates a database containing flight operations data provided by Brazilian National Civil Aviation Agency (ANAC) [10] and airport weather data provided by Weather Underground [11]. From this dataset, many data preprocessing methods were applied in combination with different machine learning classification models. Their performance evaluation regarding delay prediction was analyzed.

This paper contributes by exploring a broader spectrum of data preprocessing methods for building machine learning models. Although flight delay prediction is an open problem, our results indicate that data preprocessing methods that target the problem of unbalanced distribution of the classes for delays outperform the other preprocessing methods. Additionally, we also contributed by exploring the Brazilian flight systems, since no other work was observed targeting predicting flight delays in this continental country [12].

Besides this introduction, this paper is organized as follows. Sections II and III present the general background for data preprocessing and machine learning methods. Section IV presents related work, whereas Section V discusses the methodology used for our exploratory analysis. Section VI analyzes the experimental evaluation. Finally, Section VII concludes and point out future work.

II. Data Preprocessing

The data preprocessing activities allow for the input dataset to be prepared for the following data mining activities. Especially when it comes to classification, they also help in increasing both accuracy and performance of the machine learning methods. Among the main activities of data preprocessing we can highlight: Data Integration & Cleaning; (A) Data Transformation (DT), (B) Data Reduction (DR), and (C)

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

Knowledge-Based Systems 164 (2019) 274–291
Contents lists available at ScienceDirect
Knowledge-Based Systems
Journal homepage: www.elsevier.com/locate/knosys

Nonstationary time series transformation methods: An experimental review

Rebecca Salles [a], Kele Belloze [a], Fabio Porto [b], Pedro H. Gonzalez [a], Eduardo Ogasawara [a,*]
[a] Federal Center for Technological Education of Rio de Janeiro (CEFET/RJ), Brazil
[b] National Laboratory for Scientific Computing (LNCC), Brazil

ARTICLE INFO
Article history:
Received 2 June 2018
Received in revised form 24 October 2018
Accepted 20 October 2018
Available online 3 November 2018
Keywords:
Nonstationarity
Time series
Transformation methods
Prediction models
Review

ABSTRACT
Data preprocessing is a crucial step for mining and learning from data, and one of its primary activities is the transformation of data. This activity is very important in the context of time series prediction since most time series models assume the property of stationarity, i.e., statistical properties do not change over time, which in practice is the exception and not the rule in most real datasets. There are several transformation methods designed to treat nonstationarity in time series. However, the choice of a transformation that is appropriate to the adopted data model and to the problem at hand is not a simple task. This paper provides a review and experimental analysis of methods for transformation of nonstationary time series. The focus of this work is to provide a background on the subject and a discussion on their advantages and limitations to the problem of time series prediction. A subset of the reviewed transformation methods is compared through an experimental evaluation using benchmark datasets from time series prediction competitions and other real macroeconomic datasets. Suitable nonstationary time series transformation methods provided improvements of more than 30% in prediction accuracy for half of the evaluated time series and improved the prediction in more than 95% for 10% of the time series. Furthermore, the adoption of a validation phase during model training enables the selection of suitable transformation methods.

1. Introduction

Adequate data preprocessing is an important activity in any application aiming at data analytics. It generally demands a long time and dedication [1,2]. The main objective of data preprocessing is ensuring the quality of data serving as input to applied learning methods and therefore avoid obtaining inaccurate and/or incorrect results and conclusions [3]. Among the activities commonly performed during preprocessing, we can list data cleaning, feature and sample selection, outlier removal, normalization, and data transformation.

The data transformation activity becomes particularly important in the context of prediction [2,4]. This context encompasses both problems of classification (prediction of discrete data) and regression (prediction of continuous data) [2,4,5]. However, henceforth we focus on the problem of predicting numeric time series data through regression. For simplicity, we may refer to prediction and regression interchangeably in this paper.

Although a great variety of time series prediction methods exists in literature [6], many of these methods and the majority of works that handle time series assume that the available time series is stationary [7]. In a stationary time series, statistical properties, such as mean, variance and covariance, remain constant over time and in any sample of data [7,8]. However, in practice, we observe that such properties are not constant in the majority of real-world time series, especially in socioeconomics [9], where many of them are nonstationary. Thus, when observed the presence of nonstationarity in a time series, it is a usual approach to search for ways to transform them to achieve stationarity so that the known time series prediction methods can be applied.

There exist several transformation methods in literature for coping with nonstationarity in times series. However, the choice for an adequate method to a particular time series application is not a simple task. The analysis of their features and expected advantages is crucial. Some of the features that should be considered are their initial data assumptions (including different kinds of nonstationarity, linearity, and seasonality) and their intrinsic properties (mathematical transformation or computational algorithm). In this context, a thorough overview of different transformation methods for handling nonstationary time series and their respective features becomes particularly important.

Although the literature on nonstationarity in time series is gaining attention, not many authors focus on reviewing transformation methods for nonstationarity treatment [10,6]. Yang and Zurbenko

* Corresponding author.
E-mail address: eogasawara@ieee.org (E. Ogasawara).

https://doi.org/10.1016/j.knosys.2018.10.041
0950-7051/0 © 2018 Elsevier B.V. All rights reserved.

International Journal of Agricultural and Environmental Information Systems, 4(2), 23-36, April-June 2013 23

A Forecasting Method for Fertilizers Consumption in Brazil

Eduardo Ogasawara, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Daniel de Oliveira, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Fabio Paschoal Junior, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Rafael Castaneda, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Myrna Amorim, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Renato Mauro, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Jorge Soares, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
João Quadros, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil
Eduardo Bezerra, CEFET-Federal Center of Technological Education Celso Suckow da Fonseca, Rio de Janeiro, Brazil

ABSTRACT
Tracking information about fertilizers consumption in the world is very important since they are used to produce agriculture commodities. Brazil consumes a large amount of fertilizers due to its large-scale agriculture field. Most of these fertilizers are currently imported. The analysis of consumption of major fertilizers, such as Nitrogen-Phosphorus-Potassium (NPK), Sulfur, Phosphate Rock, Potash, and Nitrogen become critical for long-term government decisions. In this paper we present a method for fertilizers consumption forecasting based on both Autoregressive Integrated Moving Average (ARIMA) and logistic function models. Our method was used to forecast fertilizers consumption in Brazil for the next 20 years considering different economic growth for the entire country.

Keywords: *Autoregressive Integrated Moving Average (ARIMA), Fertilizers Consumption, Forecast, Gross Domestic Product (GDP), Logistic Function, Population*

DOI: 10.4018/jaeis.2013040103

Temas de Pesquisa: Workflows for data analytics

- Workflows
  - Modelagem: Atividades e dependências
  - Datacêntrica
- DISC (Data intensive scalable computing)
- Álgebra de workflow

An Algebraic Approach for Data-Centric Scientific Workflows*

Eduardo Ogasawara[1,2], Jonas Dias[1], Daniel de Oliveira[1], Fábio Porto[3], Patrick Valduriez[4], Marta Mattoso[1]

[1]COPPE/UFRJ Rio de Janeiro, Brazil; [2]CEFET/RJ Rio de Janeiro, Brazil; [3]LNCC Petrópolis, Brazil; [4]INRIA & LIRMM Montpellier, France

{ogasawara,jonasdias,danielc,marta}@cos.ufrj.br fporto@lncc.br Patrick.Valduriez@inria.fr

ABSTRACT

Scientific workflows have emerged as a basic abstraction for structuring and executing scientific experiments in computational environments. In many situations, these workflows are computationally and data intensive, thus requiring execution in large-scale parallel computers. However, parallelization of scientific workflows remains low-level, ad-hoc and labor-intensive, which makes it hard to exploit optimization opportunities. To address this problem, we propose an algebraic approach (inspired by relational algebra) and a parallel execution model that enable automatic optimization of scientific workflows. We conducted a thorough validation of our approach using both a real oil exploitation application and synthetic data scenarios. The experiments were run in Chiron, a data-centric scientific workflow engine implemented to support our algebraic approach. Our experiments demonstrate performance improvements of up to 226% compared to an ad-hoc workflow implementation.

1. INTRODUCTION

Many scientific experiments are based on complex computer simulations that consume and produce very large datasets and allocate huge amounts of computational resources. As the complexity of the experiments growths, running simulations becomes a challenge. To help scientists in managing resources involved in large-scale in-silico simulations, scientific workflows are gaining much interest. A workflow can be defined as a model of a process, which consists in a series of activities and its dependencies [1]. Workflows have been used primarily in business data processing. A data-centric scientific workflow, for short scientific workflow, structures the processing of a scientific simulation as a graph of activities, in which nodes correspond to data processing activities and edges represent the dataflow between them. Workflow activities are associated to scientific programs that prepare, process and analyze data.

Scientific Workflow Management Systems (SWfMS) [2] are software systems that support the definition, execution and monitoring of scientific workflows. Various SWfMS have been proposed (e.g. VisTrails, Kepler, Taverna, Pegasus, Swift and Triana). Each of them has its own language [2] and focuses on different aspects, such as parallel execution, semantic support, domain specific characteristics and management of provenance data.

Although some SWfMS focus on parallel execution, parallelizing large-scale simulations are still hard, ad-hoc and labor-intensive. Workflow developers (and scientists) need to decide on the ordering, dependencies, and the parallelization strategies. These decisions tighten parallelization opportunities, which may yield to miss important optimization opportunities. Let us illustrate the problem with a critical application we are addressing with Petrobras, Brazil's giant oil company. We will use this example consistently in the rest of the paper.

1.1 Motivating Example: RFA application

To illustrate the problem of optimizing data-centric scientific workflows, let us consider the following motivating workflow scenario from oil exploitation. A major function of an ultra-deep water oil exploitation system is pumping oil from thousand meters up to the surface through tubular structures, called risers. Maintaining and repairing risers under deep water is difficult, costly and critical for the environment (e.g. to prevent oil spill). Understanding the dynamic behavior of each riser and its life expectancy is critical for Petrobras. Thus, scientists must predict risers fatigue based on complex scientific models and observed data collected from risers. As shown in Figure 1, performing Risers Fatigue Analysis (RFA) requires a complex workflow of data-intensive activities that may take a very long time to compute. A typical riser's fatigue analysis workflow [3] consumes several input files containing riser information, such as finite element meshes, wind, waves, sea currents, case studies, and produces result analysis files to be further studied by the scientists. Such workflow can be modeled as a Directed Acyclic Graph (DAG) of activities with edges for establishing the dataflow. In Figure 1, each rectangle indicates an activity, solid lines represent input and output parameters that are transferred between activities, and dashed lines represent input and output parameters shared through data files. For simplicity, we assume that all data is shared among all activities, i.e., with a shared disk storage. For a single riser's fatigue analysis, the workflow can have as many as 2,000 input files (about 1GB) and produces about 6,500 files (about 22GB) with nine activities including:

* Work partially sponsored by CAPES, CNPq and INRIA (Dathaga and Sarava projects).

Rumo à Integração da Álgebra de Workflows com o Processamento de Consulta Relacional*

João Antonio Ferreira[1], Jorge Soares[1], Fabio Porto[2], Esther Pacitti[3], Rafaelli Coutinho[1], Eduardo Ogasawara[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

[2]LNCC - Laboratório Nacional de Computação Científica

[3]Inria & University of Montpellier

joao.parana@acm.org, jorge@eic.cefet-rj.br, fporto@lncc.br, Esther.Pacitti@inria.fr, rafaelli.coutinho@cefet-rj.br, eogasawara@ieee.org

Abstract. *Workflows emerged as a basic abstraction for structuring data analysis experiments in the current Data Intensive Scalable Computing (DISC) scenario. In many situations, these workflows are intensive, either computationally or in relation to data management, requiring execution in high-performance processing environments. However, parallelizing the execution of workflows commonly requires laborious programming, in an ad hoc manner and in a low level of abstraction, which makes it difficult to explore optimization opportunities. Some algebraic approaches have been developed to mitigate such limitation. This work moves in the direction converging the workflow algebra with relational query processing.*

Resumo. *Os workflows emergiram como uma abstração básica para estruturar experimentos de análise de dados no atual cenário de DISC (*Data Intensive Scalable Computing*). Em muitas situações, esses workflows são intensivos, seja computacionalmente ou em relação à manipulação de dados, exigindo a execução em ambientes de processamento de alto desempenho. Entretanto, paralelizar a execução de workflows comumente requer programação trabalhosa, de modo ad hoc e em baixo nível de abstração, o que torna difícil a exploração das oportunidades de otimização. Algumas abordagens algébricas foram desenvolvidas visando mitigar tal limitação. Este trabalho caminha na direção de convergir a álgebra de workflows com o processamento de consultas relacionais.*

1. Contexto

Apesar de alguns sistemas de workflows possuírem recursos para execução paralela, paralelizar um workflow de larga escala é uma tarefa difícil, *ad hoc* e trabalhosa. Na maioria das soluções existentes, cabe aos usuários dos sistemas decidirem a ordem e as dependências entre as atividades além das estratégias de paralelização. Estas decisões, em muitos casos, restringem as oportunidades de otimização da execução do workflow que poderiam levar a melhorias significativas de desempenho [Ogasawara et al., 2011], principalmente

*Os autores agradecem à FAPERJ, à CAPES e ao CNPq pelo financiamento parcial do projeto.

Exploring Machine Learning Methods for the Star/Galaxy Separation Problem

Eduardo Machado, Marcelo Serqueira, Eduardo Ogasawara, CEFET/RJ, eogasawara@ieee.org

Ricardo Ogando, Marcio A. G. Maia, Luiz Nicolaci da Costa, Observatório Nacional, LIneA, {ogando, ldacosta, maia}@linea.gov.br

Riccardo Campisano, Gustavo Paiva Guedes, Eduardo Bezerra, CEFET/RJ, {gguedes, ebezerra}@cefet-rj.br

Abstract—**For recent or planned deep astronomical surveys, it is important to tell stars and galaxies apart, a task known as Star/Galaxy Separation Problem (SGSP). At faint magnitudes, the separation between points and extended sources is fuzzy, which makes SGSP a hard task. This problem is even harder for large surveys like Dark Energy Survey (DES) and, in a near future, the Large Synoptic Survey Telescope (LSST) due to their large data volume. Hence, the search for classification methods that are both accurate and efficient is highly relevant. In this work, we present a comparative analysis of several machine learning methods targeted at solving the SGSP at faint magnitudes. In order to train the classification models, the COSMOS survey was used. We use machine learning methods as distinct as artificial neural networks, k nearest-neighbor, Support Vector Machines, Random Forests and Naive Bayes. The exploratory process was modeled as data centric workflow. The workflow was implemented on top of Hadoop framework and was used to find the best parameter values for each classification method we considered, of which neural networks and random forest present superior performance.**

I. Introduction

A lot has changed since the first astronomical surveys [1] when, along with the well-known stars and planets, a new class of extended sources, the nebulae, was discovered. It was only in the beginning of the XX century that this new class was later associated to extra-galactic sources, the galaxies. This distinction between extended and point sources was enough to tell them apart until recently.

As astronomical surveys pushed the boundaries of the observed universe, reaching fainter magnitude limits and challenging the spatial resolution of ground-based telescopes, this distinction between stars and galaxies became fuzzier, as one just can not spatially resolve distant galaxies, confusing them with stars (or QSOs).

One of the main issues that affect the spatial resolution of ground-based telescopes is the blur caused by atmosphere turbulence, or seeing. Removing the atmosphere from the line of sight using space telescopes, like Hubble Space Telescope (HST [2]), one can resolve distant galaxies that would be otherwise indistinguishable from stars when observed from Earth.

The discovery of the accelerated expansion of the universe [3], caused by an unknown component generally called Dark Energy, started a race to uncover its nature. Several large and deep photometric surveys are ongoing (DES [4]) or under construction (LSST [5]) with the main goal of studying dark energy. In order to do that, cosmological probes such as large scale structure, weak lensing, and galaxy clusters, require a very good distinction between stars and galaxies. While surveys like SDSS [6] relied successfully on morphological classification on its data releases, there is an urge for new techniques when dealing with deeper surveys. Many works explore machine learning techniques [7], [8], specially using crispy HST images from COSMOS survey [2] as training set.

As the amount of data collected in deeper surveys grows, the demand for fast and automatic star/galaxy separation methods is paramount. Many studies exist that use one or two classification methods. In particular, artificial neural networks have been a common tool [7], and, more recently, support vector machines [8]. On the other hand, it is well-known that no algorithm performs well on all possible classification problems [9]. To the best of our knowledge, the literature for star/galaxy separation problem lacks (1) a more extensive experimental analysis of a broader spectrum of machine learning methods, specially near the faint/unresolvable end of magnitude/shape distribution, and (2) a disciplined approach to fine-tune the parameters of such methods.

In this paper, we explore the application of several machine learning methods to the star/galaxy separation problem, namely, neural networks, SVM, random forests, k-nearest neighbors and Naive Bayes. To train and validate these methods, we use COSMOS survey [2], which consists of a catalog of more than 500,000 objects and 90 attributes. The exploratory process was modeled as data centric workflow and implemented on top of Hadoop framework. Our approach takes into account preprocessing activities and parameters exploration that are not commonly used in this scenario. Our experiments show that neural networks and random forest present superior performance than the remaining methods. The result is important because random forest is not among the most used methods in star/galaxy separation problem, which reinforces the need for such exploration.

This work is organized in the following sections. Section II presents an overview on Astronomy surveys and related concepts. Section III provides a brief description of the preprocessing techniques whereas Section IV presents the

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

Temas de Pesquisa: Pattern mining

- Mineração de sequências
- Descoberta de motifs
- Detecção de eventos, anomalias, outliers

Transportation Research Part E 95 (2016) 282–298

Contents lists available at ScienceDirect

Transportation Research Part E

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

An analysis of Brazilian flight delays based on frequent patterns

Alice Sternberg [a], Diego Carvalho [a], Leonardo Murta [c], Jorge Soares [a,b], Eduardo Ogasawara [a,*]

[a] Federal Center of Technological Education of Rio de Janeiro - CEFET/RJ, Brazil
[b] State University of Rio de Janeiro - UERJ, Brazil
[c] Fluminense Federal University - UFF, Brazil

ARTICLE INFO

Article history:
Received 23 March 2016
Received in revised form 31 August 2016
Accepted 30 September 2016

Keywords:
Flight delays
Association rules
Commercial aviation
Delay reasons

ABSTRACT

In this paper we applied data indexing techniques combined with association rules to unveil hidden patterns of flight delays. Considering Brazilian flight data and guided by six research questions related to causes, moments, differences, and relationships between airports and airlines, we evaluated and quantified all attributes that may lead to delays, showing not only the main patterns, but also their chances of occurrence in the entire network, in each airport and airline. We observed that Brazilian flight system has difficulties to recover from previous delays and when operating under adverse meteorological conditions, delays occurrences may increase up to 216%.

1. Introduction

Delays are one of the greatest challenges to transportation systems. Notably, in commercial aviation, delay is usually defined by the difference between scheduled and real times of departure or arrival (Wieland, 1997). Despite some differences in tolerance thresholds for delays, country regulatory authorities usually monitor delays levels through several indicators. In 2014, 16.6% of flights delayed by more than 15 min in Europe and 24.7% in the United States. In Brazil, 19.1% of domestic flights were canceled or suffered delays greater than 30 min (EUROCONTROL, 2015; The Unites States Department of Transportation, 2015; ANAC, 2015b).

Flight delays impact passengers, airlines, and airports, especially increasing trip and operations costs. Given the uncertainty of their occurrence, passengers usually plan to travel earlier to ensure their arrival on time. On the other hand, airlines may have to pay penalties, fines, or incur extra cost, such as crew reschedules and aircraft retentions in airports (Britto et al., 2012). Moreover, delays are also related to environmental damages, since they may increase fuel consumption and gas emissions (Pejovic et al., 2009; Ryerson et al., 2014; Simić and Babić, 2015).

Delays also affect the airlines marketing strategies, since the loyalty of customers are motivated by punctual performances (Vlachos and Lin, 2014). Furthermore, delay levels are not only related to operational and economic choices of an airline (such as aircraft sizes, flight frequencies, and fares), but also with complaints about airline service (Bhadra, 2009; Pai, 2010; Zou and Hansen, 2014). In this context, understanding the reasons for flight delays occurrences can direct public and private investments in air transportation systems, improve tactical and operational decisions of airports and airlines managers, and warn passengers, so they can rearrange their plans (Marsden, 2002; Lv and Wang, 2009).

Every moment, a massive amount of data from commercial aviation is collected and stored in public and private databases. Seeking to understand the air transportation ecosystem, domain analysts and data scientists are intensifying the usage

* Corresponding author.
E-mail address: eogasawara@ieee.org (E. Ogasawara).

http://dx.doi.org/10.1016/j.tre.2016.09.013
1366-5545/

Discovering Tight Space-Time Sequences

Riccardo Campisano[1], Heraldo Borges[1], Fabio Porto[2], Fabio Perosi[3], Esther Pacitti[4], Florent Masseglia[4], and Eduardo Ogasawara[1](✉)

[1] CEFET/RJ - Federal Center for Technological Education of Rio de Janeiro, Rio de Janeiro, Brazil
eogasawara@ieee.org
[2] LNCC - National Laboratory of Scientific Computing, Petrópolis, Brazil
[3] UFRJ - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[4] Inria and University of Montpellier, Montpellier, France

Abstract. The problem of discovering spatiotemporal sequential patterns affects a broad range of applications. Many initiatives find sequences constrained by space and time. This paper addresses an appealing new challenge for this domain: find tight space-time sequences, *i.e.*, find within the same process: (i) frequent sequences constrained in space and time that may not be frequent in the entire dataset and (ii) the time interval and space range where these sequences are frequent. The discovery of such patterns along with their constraints may lead to extract valuable knowledge that can remain hidden using traditional methods since their support is extremely low over the entire dataset. We introduce a new *Spatio-Temporal Sequence Miner* (*STSM*) algorithm to discover tight space-time sequences. We evaluate *STSM* using a proof of concept use case. When compared with general spatial-time sequence mining algorithms (*GSTSM*), *STSM* allows for new insights by detecting maximal space-time areas where each pattern is frequent. To the best of our knowledge, this is the first solution to tackle the problem of identifying tight space-time sequences.

1 Introduction

Space and time are pervasive in our day-to-day lives. As many datasets that include both time and space data are becoming available, new opportunities to discover interesting spatiotemporal patterns arise. An event may be classified as an occurrence of a phenomenon in a given space and time. A spatiotemporal sequential pattern is a sequence of events that are constrained in space and time [7]. Due to that, spatiotemporal sequence mining is gaining attention [11,12].

In this work, we investigate a new problem related to spatiotemporal pattern identification. We are interested in finding tight space-time sequences, *i.e.*, sequences that are constrained in space and time that may not be frequent in the entire dataset but are frequent inside a time interval and space range (spatiotemporal blocks). The primary challenge is to discover these blocks and the frequent sequences they contain. Solving this problem has a valuable impact on many applications.

C. Ordonez and L. Bellatreche (Eds.): DaWaK 2018, LNCS 11031, pp. 247–257, 2018.
https://doi.org/10.1007/978-3-319-98539-8_19

2018 SBC 33rd Brazilian Symposium on Databases (SBBD) August 25-28, 2018 - Rio de Janeiro, RJ, Brazil

Detecção de Anomalias Frequentes no Transporte Rodoviário Urbano*

Ana Beatriz Cruz[1], João Ferreira[1], Diego Carvalho[1], Eduardo Mendes[4] Esther Pacitti[5], Rafaelli Coutinho[1], Fabio Porto[2], Eduardo Ogasawara[1]

[1] CEFET/RJ
[2]LNCC - DEXL Lab
[3]FGV
[4]Inria & University of Montpellier

anacruz@acm.org, joao.parana@acm.org, d.carvalho@ieee.org, eduardo.mendes@fgv.br, renato.souza@fgv.br, Esther.Pacitti@inria.fr, rafaelli.coutinho@cefet-rj.br, fporto@lncc.br, eogasawara@ieee.org

Abstract. *The growth of urban population and, consequently, the number of vehicles causes the increase of traffic jams and emission of polluting gases. In this context, we observe the intensification of papers that aim to identify bottlenecks and their causes. These papers propose methodologies that use trajectory data model and aim to explain systemic behaviors. This article proposes the identification and classification of anomalies in the urban road transport system from space-time aggregations to permanent objects. The methodology consists of pre-processing of data, identification of anomalies, identification, and classification of frequent patterns. Through it, we can identify the systemic and specific behaviors on the urban transit of Rio de Janeiro.*

Resumo. *O crescimento da população urbana e, consequentemente, do número de veículos provoca o aumento de engarrafamentos e da emissão de gases poluentes. Nesse contexto, observa-se a intensificação de pesquisas que buscam identificar engarrafamentos e suas causas. Estas pesquisas propõem metodologias que usam modelo de dados de trajetória e visam explicar comportamentos sistêmicos. Este artigo propõe a identificação e a classificação de anomalias no sistema de transporte rodoviário urbano a partir de agregações espaço-temporais a objetos permanentes. A metodologia consiste do pré-processamento dos dados, identificação de anomalias, identificação e classificação de padrões frequentes. Por meio dela, é possível identificar comportamentos sistêmicos e pontuais do trânsito urbano do Rio de Janeiro.*

1. Introdução

Em 2007, pela primeira vez existiam mais pessoas vivendo em áreas urbanas do que em zonas rurais, resultado de uma urbanização expressiva que se impulsionou desde a década

*Os autores agradecem à FAPERJ, à CAPES e ao CNPq pelo financiamento parcial do projeto.

271

Colaboração

- Institutos de Pesquisa
  - LNCC
    - Fabio Porto, Artur Ziviani, Antônio Tadeu A. Gomes
  - Fiocruz
    - Marcel Pedroso, Cristiano Boccolini, Christovam Barcellos
- Academia
  - CEFET/RJ
    - A maioria dos docentes do PPCIC/PPPRO
  - COPPE/UFRJ
    - Marta Mattoso, Geraldo Zimbrão, Geraldo Xexéo
  - UFF
    - Daniel Oliveira, Leonardo Murta, Vanessa Braganholo
- Internacionais
  - INRIA / University of Montpellier
    - Patrick Valduriez, Esther Pacitti, Florent Masseglia, Reza Akbarinia