ARTIFICIAL INTELLIGENCE AND DATA STREAM MINING

ALBERT BIFET

15 FEVRIER 2018

#Futur&Ruptures
SOMMAIRE

1. ARTIFICIAL INTELLIGENCE
2. ARTIFICIAL INTELLIGENCE CHALLENGES
3. MACHINE LEARNING FOR DATA STREAMS
4. OPEN SOURCE TOOLS
5. SUMMARY
ARTIFICIAL INTELLIGENCE
Artificial Intelligence is the new Electricity

- Big Data
- Internet of Things
- Data Science
- Artificial Intelligence
What is AI?

- Artificial intelligence (AI) is an area of computer science that emphasizes the creation of intelligent machines.”
ARTIFICIAL INTELLIGENCE CHALLENGES
Artificial Intelligence

47% Of All Jobs Will Be Automated By 2034, And ‘No Government Is Prepared’ Says Economist

Michael Rundle
HuffPost UK Technology Editor
Artificial Intelligence

Is Your Job At Risk?²

- Doctoral or Professional Degree
- Master’s
- Bachelor’s
- Associate’s
- Postsecondary Nondegree Award
- Some College
- High School Diploma or Equivalent
- No Formal Education Credential

Search by occupation:

The best paid, least vulnerable occupations are doctors, dentists and CEOs.

The best paid, most vulnerable occupations include accountants, benefits managers, credit analysts, and various insurance professionals.

Some lower-wage jobs with higher-education requirements are less likely to be automated.

Low-paid occupations also tend to be most at-risk. They include cashiers, drivers, and food service workers.

Average annual wage

DATA: FREY & OSBORN, BUREAU OF LABOR STATISTICS
Laurent Alexandre : « L'Europe a complètement perdu la bataille de l'IA »

DAVID BARROUX | BENOIT GEORGES | 17/11/2017

Laurent Alexandre, chirurgien urologue, cofondateur du site web Doctissimo.fr, auteur du livre "La défaite du cancer" (éditions JC Lattes).

+ VIEDO. Passionné, sûr de lui et volontiers provocateur, Laurent Alexandre occupe une place à part dans le petit monde des futuroloques français. Il fut l'un des premiers à alerter sur la puissance
Big Data

• GAFAM: Google, Apple, Facebook, Amazon, Microsoft

• Personal Information
  • Google, Facebook, Twitter, Linkedin,..
  • All personal communications in Europe are managed by non-European companies
Pour Cédric Villani, "la France n'a pas perdu la guerre de l'IA"
According to Nikola Kasabov, AI systems should exhibit the following characteristics:

- Accommodate new problem solving rules **incrementally**
- **Adapt online and in real time**
- Are able to **analyze itself** in terms of behavior, error and success.
- Learn and improve through interaction with the environment (embodiment)
- Learn quickly from large amounts of data (**Big Data**)
- Have memory-based exemplar storage and retrieval capacities
- Have parameters to represent short and long term memory, age, forgetting, etc.
La vision de la France doit donc consister à développer simultanément une IA plus verte et une IA au service de la transition écologique.
Machine Learning

• **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

• **Machine learning** focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.
Analytic Standard Approach

Finite training sets
Static models
Data Stream Approach

Infinite training sets
Dynamic models
Importance of Online Learning

- As spam trends change, it is important to retrain the model with newly judged data.

*Previously tested using news comment in Y! Inc.*

- Over 29 days period, you can see degradation in performance of base model (w/o active learning) VS online model (AUC stands for Area Under Curve).

Pain Points

- Need to **retrain**!
- Things change over time
- How often?
- Data unused until next update!
- Value of data wasted
IoT Stream Mining

- Maintain models online
  - Incorporate data on the fly
  - Unbounded training sets
  - Resource efficient
  - Detect changes and adapts
  - Dynamic models
Approximation Algorithms

• General idea, good for streaming algorithms

• Small error $\varepsilon$ with high probability $1-\delta$
  
  • True hypothesis $H$, and learned hypothesis $\hat{H}$
  
  • $\Pr[ |H - \hat{H}| < \varepsilon|H| ] > 1-\delta$
Hoeffding Adaptive Tree

- Replace frequency counters by estimators
  - No need for window of instances
  - Sufficient statistics kept by estimators separately
- Parameter-free change detector + estimator with theoretical guarantees for subtree swap (ADWIN)
  - Keeps sliding window consistent with “no-change hypothesis”
ADWIN

An adaptive sliding window whose size is recomputed online according to the rate of change observed.

Problem
Given an input sequence $x_1, x_2, \ldots, x_t, \ldots$ we want to output
- a prediction $\hat{x}_{t+1}$ minimizing prediction error:
  $$|\hat{x}_{t+1} - x_{t+1}|$$
- an alert if change is detected
ADWIN

Optimal Change Detector and Predictor
- High accuracy
- Fast detection of change
- Low false positives and false negatives ratios
- Low computational cost: minimum space and time needed

ADWIN
- Theoretical guarantees
- No parameters needed
ADWIN

Theorem
At every time step we have:

1. (False positive rate bound). If $\mu_t$ remains constant within $W$, the probability that ADWIN shrinks the window at this step is at most $\delta$.

2. (False negative rate bound). Suppose that for some partition of $W$ in two parts $W_0 W_1$ (where $W_1$ contains the most recent items) we have $|\mu_{W_0} - \mu_{W_1}| > 2\varepsilon_c$. Then with probability $1 - \delta$ ADWIN shrinks $W$ to $W_1$, or shorter.

ADWIN tunes itself to the data stream at hand, with no need for the user to hardwire or precompute parameters.
ADWIN

**Classification**
- Adaptive Naive Bayes (Bifet et al. 2007)
- Decision Trees: Hoeffding Adaptive Trees (Bifet et al. 2009)
- ADWIN Bagging (Bifet et al. 2009)
- Leveraging Bagging (Bifet et al. 2010)
- Stacking of Restricted Hoeffding Trees (Bifet et al. 2012)
- Multilabel Classification (Read et al. 2012)
- Adaptive kNN (Bifet et al. 2013)
- Random Forests (Marron et al. 2014)

**Frequent Pattern Mining**
- Frequent Closed Tree Mining (Bifet et al. 2008)
- Frequent Closed Graph Mining (Bifet et al. 2011)
Adaptive Random Forest

• Why Random Forests?
  • Off-the-shelf learner
  • Good learning performance

Adaptive random forests for evolving data stream classification.

Gomes, H M; Bifet, A; Read, J; Barddal, J P; Enembreck, F; Pfharinger, B; Holmes, G; Abdessalem, T.


• Based on the original Random Forest by Breiman
OPEN SOURCE TOOLS
MOA

• Massive Online Analysis is a framework for online learning from data streams.

• It is closely related to WEKA

• It includes a collection of offline and online as well as tools for evaluation:
  • classification, regression
  • clustering, frequent pattern mining

• Easy to extend, design and run experiments
APACHE SAMOA

G. De Francisci Morales, A. Bifet: “SAMOA: Scalable Advanced Massive Online Analysis”. JMLR (2014)
Creating a Flink Adapter on Apache SAMOA

Apache Scalable Advanced Massive Online Analysis (SAMOA) is a platform for mining data streams with the use of distributed streaming Machine Learning algorithms, which can run on top of different Data Stream Processing Engines (DSPEs).

As depicted in Figure 20, Apache SAMOA offers the abstractions and APIs for developing new distributed ML algorithms to enrich the existing library of state-of-the-art algorithms [27, 28]. Moreover, SAMOA provides the possibility of integrating new DSPEs, allowing in that way the ML programmers to implement an algorithm once and run it in different DSPEs [28].

An adapter for integrating Apache Flink into Apache SAMOA was implemented in scope of this master thesis, with the main parts of its implementation being addressed in this section. With the use of our adapter, ML algorithms can be executed on top of Apache Flink. The implemented adapter will be used for the evaluation of the ML pipelines and HT algorithm variations.

5.1 Apache SAMOA Abstractions

Apache SAMOA offers a number of abstractions which allow users to implement any distributed streaming ML algorithms in a platform independent way. The most important abstractions of Apache SAMOA are presented below [27, 28].
http://huawei-noah/github.io/streamDM

StreamDM

streamDM: Data Mining for Spark Streaming
scikit-multiflow

A multi-output/multi-label and stream data framework. Inspired by MOA and MEKA, following scikit-learn philosophy.

[GitHub Repository]

Photo credit: freepik

Mainly supported by:

© 2018 scikit-multiflow. Website powered by Jekyll & Minimal Mistakes.
SUMMARY
IoT: sensors and actuators connected by networks to computing systems.
- Gartner predicts 20.8 billion IoT devices by 2020.
- IDC projects 32 billion IoT devices by 2020
IoT versus Big Data
# Applications IoT Analytics

## IoT Segment

<table>
<thead>
<tr>
<th>Segment</th>
<th>Global share of IoT projects&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Americas</td>
</tr>
<tr>
<td>Connected Industry</td>
<td>22%</td>
<td>43%</td>
</tr>
<tr>
<td>Smart City</td>
<td>20%</td>
<td>31%</td>
</tr>
<tr>
<td>Smart Energy</td>
<td>13%</td>
<td>49%</td>
</tr>
<tr>
<td>Connected Car</td>
<td>13%</td>
<td>43%</td>
</tr>
<tr>
<td>Other</td>
<td>8%</td>
<td>46%</td>
</tr>
<tr>
<td>Smart Agriculture</td>
<td>6%</td>
<td>48%</td>
</tr>
<tr>
<td>Connected Building</td>
<td>5%</td>
<td>48%</td>
</tr>
<tr>
<td>Connected Health</td>
<td>5%</td>
<td>61%</td>
</tr>
<tr>
<td>Smart Retail</td>
<td>4%</td>
<td>52%</td>
</tr>
<tr>
<td>Smart Supply Chain</td>
<td>4%</td>
<td>57%</td>
</tr>
</tbody>
</table>

N = 640 global, publicly announced IoT projects

1. Based on 640 publicly known enterprise IoT projects (not including consumer IoT projects e.g. Wearables, Smart Home) 2. Trend based on IoT Analytics’ Q3/2016 IoT Deployment Statistics Tracker 3. Not including Consumer Smart Home Solutions 4. Source: IoT Analytics 2016 Global overview of IoT enterprise IoT use cases (August 2016)
IOT AND INDUSTRY 4.0

- Interoperability: IoT
- Information transparency: virtual copy of the physical world
- Technical assistance: support human decisions
- Decentralized decisions: make decisions on their own
Overview

Today many information sources—including sensor networks, financial markets, social networks, and healthcare monitoring—are so-called data streams, arriving sequentially and at high speed. Analysis must take place in real time, with partial data and without the capacity to store the entire data set. This book presents algorithms and techniques used in data stream mining and real-time analytics. Taking a hands-on approach, the book demonstrates the techniques using MOA (Massive Online Analysis), a popular, freely available open-source software framework, allowing readers to try out the techniques after reading the explanations.

The book first offers a brief introduction to the topic, covering big data mining, basic methodologies for mining data streams, and a simple example of MOA. More detailed discussions follow, with chapters on sketching techniques, change, classification, ensemble methods, regression, clustering, and frequent pattern mining. Most of these chapters include exercises, an MOA-based lab session, or both. Finally, the book discusses the MOA software, covering the MOA graphical user interface, the command line, use of its API, and the development of new methods within MOA. The book will be an essential reference for readers who want to use data stream mining as a tool, researchers in innovation or data stream mining, and programmers who want to create new algorithms for MOA.
Data, Intelligence and Graphs (DIG)

DIG is a group of researchers at LTCI, Télécom ParisTech who study the fundamental issues raised in data and knowledge management systems, graph mining and artificial intelligence. Research interests cover theoretical foundations of data intelligence and graph systems, practical solutions and applications, as well as cognitive aspects. The group was formerly known as DBWeb Team.

DIG has strong academic and industrial collaborations:
Data, Intelligence and Graphs (DIG)

Senior

- Talel Abdessalem, Professor
- Antoine Amarilli, Associate Professor
- Albert Bifet, Professor, Head
- Thomas Bonald, Professor
- Laurent Decreusefond, Professor
- Jean-Louis Dessalles, Associate Professor
- Yanlei Diao, Professor (co-affiliated with École polytechnique)
- Pierre Senellart, Professeur Invité
- Mauro Sozio, Associate Professor
- Fabian M. Suchanek, Professor
Jean-Louis Dessalles

Cet article est une ébauche concernant les sciences cognitives.
Vous pouvez partager vos connaissances en l'améliorant (comment ?) selon les recommandations des projets correspondants.

Pour les articles homonymes, voir Dessalles.

Jean-Louis Dessalles, né le 1er janvier 1956 à Périgueux, est un informaticien français, chercheur en Intelligence Artificielle et en Sciences cognitives, professeur à Télécom ParisTech, connu notamment pour ses contributions à la Théorie de la simplicité et pour une théorie originale sur l'origine (politique) du langage.

Sommaire

1 Biographie
2 Recherches
3 Bibliographie sélective
4 Voir aussi
   4.1 Articles connexes
   4.2 Liens externes
5 Références

Biographie
YAGO

Cet article est une ébauche concernant le Web sémantique.
Vous pouvez partager vos connaissances en l'améliorant (comment ?) — pour plus d'indications, visitez le projet Web Sémantique.

YAGO (Yet Another Great Ontology) est une base de connaissance créée par l'institut Max-Planck d'informatique à Sarrebruck. Elle est constituée à partir d'informations extraites de Wikipédia et d'autres sources.

En 2012, YAGO2, la deuxième version de YAGO, possède 10 millions d'entités avec plus de 120 millions d'informations à propos de ces entités². Les connaissances de YAGO sont extraites de Wikipédia (catégories, redirections, infoboxes), de WordNet (synsets, hyponymie), et de GeoNames³. La fiabilité de YAGO a été manuellement évaluée sur un échantillon de faits et est supérieure à 95%⁴. Afin de l'intégrer au web des données, YAGO est lié aux ontologies DBpedia⁵ et SUMO⁶.

YAGO2 est fourni dans la syntaxe Turtle et au format tsv. Des sauvegardes totales ou dumps de la base de données sont disponibles, des sauvegardes partielles thématiques ou spécialisées sont également disponible. Des requêtes peuvent être lancées à partir de
Thanks!

@abifet
ARTIFICIAL INTELLIGENCE AND DATA STREAM MINING

ALBERT BIFET

15 FEVRIER 2018

#Futur&Ruptures