Predicting Product Quality as Early as Possible in the Production Process

G. Fricout
ArcelorMittal Global R&D

Data Science Solutions
For Industrial Challenges

IDS 2017
Industrial Data Science Conference
September 5th, 2017, Dortmund, Germany
Overview

• (Very) Brief presentation of ArcelorMittal global R&D

• Data mining: why and how

• PRESED RFCS overview
  – Theoretical developments: the “shapelet” approach
  – Building an “industrial” tool
  – Knowledge management

• Conclusion and perspective
ArcelorMittal

At a glance
ArcelorMittal
Transforming tomorrow

- The world's leading steel and mining company, with around 199,000 employees in more than 60 countries
- Leader in all major global steel markets, including Automotive, Construction, Household appliances and Packaging, with leading R&D and technology
- An industrial presence in 18 countries exposes the company to all major markets, from emerging to mature
- The largest producer of steel in the EU, North and South America and Africa, a significant steel producer in the CIS region, and a growing presence in Asia, including investments in China and India
- One of the world’s largest producers of iron ore and metallurgical coal strategically positioned to serve our network of steel plants and the external global market

Underpinning all our operations is a philosophy to produce safe, sustainable steel
Global R&D
Meeting our customers’ needs
Global R&D
2016 Facts & Figures

- 1,400 full time researchers
- 2016 spending of $239m
- Broad, comprehensive portfolio and programmes addressing business needs
- Worldwide network of laboratories: 12 labs in Europe and Americas

Budget spending by focus area:
- Exploratory 6%
- Process 38%
- Product & Application 56%

Product research (%):
- Automotive
- Plates, Tubes & HR for Energy
- General Industry*
- Construction
- Other

*Packaging, appliance, metal processing, electrical steels

R&D effort fully aligned with Group strategy: geography, value chain, product differentiation
Data mining: why and how?

Lightweight, ...

strong design

Our constant goal
Data mining for knowledge discovery

Defect occurrence

Data Mining, statistical modeling:
- Extract, Transform and Prepare Data
- Identify the most important parameters
- Construct a statistical Prediction Model

➤ Identify (and understand) issues as early as possible in the production chain

When process specialists do not have an immediate answer…
Possible improvements in statistical modeling?

- Better models?
  - Robust statistical methods already used:
    - Variable selections, Bayesian Network, Decision Tree, NN...

- Or richer data?
  - Industry 4.0 provides data!
  - Feature extraction from sensor “profiles”

⇒ Theoretical complexity
⇒ “big data” volume

The example of defect prediction model:

- Process “rules” : close to random
- With “data mining methodology”:
  - simple model: +10% performance
  - Complex model: ~ similar

- Basic test with profile information:
  +6% performance
More is expected!
Scientific problem:

- Build robust statistical models:
  - With « vector of variables » …
  - … AND a set of « curves »
  - In order to predict/explain defect occurrence (for instance)
  - As « interpretable » as possible

- Key challenges:
  - Multivariate time series with:
    - variable lengths, variable scales and frequencies
  - Approximate synchronization
  - Complexity and computing time

<table>
<thead>
<tr>
<th>Coil</th>
<th>V1</th>
<th>V2</th>
<th>…</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>OK</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>KO</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>4</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>5</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>6</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>7</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>8</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>OK</td>
</tr>
<tr>
<td>9</td>
<td>...</td>
<td>...</td>
<td>…</td>
<td>OK</td>
</tr>
</tbody>
</table>
Sensor data
⇔ Time series ⇔ Signal

• Relevant information exists in the temporal evolution of the measurements
  ➢ How to discover automatically these evolutions?

• Often too simple descriptors are used
  – Global mean, standard deviations, etc.
  – Often inadequate

Measurement over time of a (physical) parameter

≠ information but same indicator
Inaccurate descriptor
PREdictive SEnsor Data mining (PRESED)
RFSR-CT-2014-00031  2014-2017

- Data collection
- Data preparation
- Data visualization

Robust algorithms
- Signal processing
- Machine learning

Hardware architecture
- Computing power
- Easy access to the data and models

Exchanges with process experts
Capitalization of knowledge obtained from statistical analysis
Highlights on algorithms

Lightweight, ...

strong design

Our constant goal
Extract relevant features from temporal data (PhD of X. Renard)

Temporal pattern discovery

- Use machine learning to identify relevant patterns → Localized/Global discriminant subsequences

X. Renard, M. Rifqi, G. Fricout, and M. Detyniecki, (Submitted to ECML/PKDD), 2016.

Define “manually” simple or complex features

- Frequency information, derivative, physical models…
- Enable prior process knowledge to be used

Complementarity

Temporal pattern discovery
Intuition on a simple example

Motif candidate

Motif occurrence is more probable for defective products

Motif unobserved on OK products
Temporal pattern discovery
Intuition on a toy example

Motif candidate

Motif occurrence is more probable for defective products

Motif discovery “supervised” by the product quality

Motif unobserved on OK products
Interest of the approach (PhD of X. Renard)

• The shapelet output is very easy to interpret and discuss with process specialists

• Methods have been optimized:
  – To select meaningful shapelets with limited computing cost
  – To handle a high number of time series
  – To select multi-scale, multi-dimensional shapelets

• Benchmarks show better performances than deep-learning techniques on several industrial data sets
Key ideas behind EAST* algorithms

*Enumerate And Select discriminant Temporal patterns

- Testing all possible sub-sequences from one data set is not tractable:
  ➔ A suitable number of candidates is randomly drawn from the training set

It works!

Relevant discriminating patterns are necessarily “recurrent”
➔ probability to draw them randomly is not too small and scales well with the number of individuals in the training set

**EAST representation**

1. Draw candidates among every subsequence
2. Minimal Euclidean distance between each drawn candidate and every time series
3. Representation as a feature space

<table>
<thead>
<tr>
<th>Min Eucl. Distance between TS &amp; Candidates</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>9.2</td>
<td>9.8</td>
<td>9.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>1.5</td>
<td>8.6</td>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>8.5</td>
<td>9.9</td>
<td>8.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#4</td>
<td>1.3</td>
<td>8.7</td>
<td>1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

4. Set of candidates selection with advanced feature selection

Typical classifier training
Scattering of annealing temperature for DP780 steel grade

• General idea
  – Understand annealing temperature scattering
    • leading to mechanical property scattering

• Data set:
  – 452* products, 169 static variables, 56 dynamic variables
    • from hot-rolling, pickling, cold-rolling, galvanizing with phasing information
  * Relatively low number of individuals (but high number of variables and data point in time-series ➔ Not completely suitable for deep-learning

• Output data to be predicted/understood:
  – Annealing temperature standard deviation
    • At the exit of the pre-heating section
    • At the soaking section
    ➔ high standard deviation ➔ bad product

2 different analyses
Performance

- **Statistical model:**
  - 44% of detection of scattered products for 10% false alarms
  ➔ 10% improvements compared to classical models

- In that case, physical interpretation is more important than absolute performance
  - Cold-rolling variables have more impacts on pre-heating temperature
  - Hot-rolling variables have more impacts on soaking temperature
Physical interpretation of shapelets

Cold rolling speed

Product length

Strong decrease in rolling speed
⇒ scattered pre-heating temperature

Cooling flow rate on ROT

Product length

“Shoulder” flow rate pattern
⇒ Limited scatter in soaking temperature
Industrial platform based on RapidMiner suite

Lightweight, ...

strong design

Our constant goal
Target: industrially usable tool

- **Design of a specific data model (MongoDB) and visualization tool (RapidMiner Server)**
  - Adapt to the steel industry specificities
  - Process usually spread over several plants
  - Different possible routes for one coil
  - Steel coils changing length and direction at each production step
    - Synchronizing the data from different areas is a real challenge

- **Use of RapidMiner Studio** to design time-series scripts based on “generic templates”.
  - Integration of theoretical developments of the project

- **Use of RapidMiner Server** to access continuously evolving data bases, visualizing the data and the model outputs
Knowledge capitalization

Lightweight, ...

strong design

Our constant goal
Knowledge capitalization

Knowledge Foundation
- Process experts
  - Defects
  - Operating condition
  - Causal relationships

Knowledge Exploitation
- Complex and systemic dependencies analysis
- Improvement proposal

Fleet-wide Capitalisation
- Situation retrieval
- Comparison

Statisticians
- Variables
- Features
- Correlation

- Focusing and guided datamining
- Model reuse
- Adaptation/Generalisation

© ArcelorMittal 2015. All rights reserved for all countries. Cannot be disclosed, used, or reproduced without prior written specific authorization of ArcelorMittal. CONFIDENTIAL - Privileged Information - ArcelorMittal proprietary information.
How to speed-up statistical analysis on new problems

- Provide “sharable” concepts from both statistics and process expertise
  - Based on ontologies

- Capitalize results of statistical analysis using this set of concepts
  - Important variables
  - Best pre-processing
  - Best features, meaningful shapelets…
  - Physical interpretation of the results

- Provide a search engine:
  - Retrieve similar situations based on physical description of the issue by plant experts

- Provide a “ready to run” data-mining script for a new analysis
  - Predefined variables, transformations, algorithms…
  - Based on KASEM software (from Predict company)
  - And RapidMiner (from RapidMiner company)
Conclusion
Conclusion

- **PRESED RFCS outputs:**
  - New efficient and interpretable algorithms to perform data-mining studies with continuous process data
  - A complete software framework, scalable to “big data”:
    - Data model (based on NoSql MongoDB data base)
    - **RapidMiner scripts** to perform all data-mining procedure
    - RapidMiner web interface for visualizing data
    - RapidMiner web interface for providing real-time dashboards and KPIs
  - A framework for knowledge capitalization
    - Based on ontologies inference engine and **KASEM software**
- Many potential areas of interest (predictive maintenance, energy…)

05/09/2017

IDS 2017 - G. Fricout
### Coil Aspect / Statistics

#### Information

The statistics below are based on the actual dataset for use case 3. Please note: The `dynamic_var_...` values only reflect the number of available values for this variable, not the values themselves.

The full statistics table can be downloaded as an Excel file by clicking on the download link on the right.

#### Distribution of Label Combinations

<table>
<thead>
<tr>
<th>Label</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>label_0</td>
<td>226</td>
</tr>
<tr>
<td>label_1</td>
<td>226</td>
</tr>
<tr>
<td>label_2</td>
<td>684</td>
</tr>
<tr>
<td>label_3</td>
<td>226</td>
</tr>
<tr>
<td>label_4</td>
<td>226</td>
</tr>
<tr>
<td>label_5</td>
<td>684</td>
</tr>
</tbody>
</table>

#### Statistics

<table>
<thead>
<tr>
<th>Label</th>
<th>Value_0</th>
<th>Value_1</th>
<th>Value_n/a</th>
</tr>
</thead>
<tbody>
<tr>
<td>label_0</td>
<td>226</td>
<td>226</td>
<td>684</td>
</tr>
<tr>
<td>label_1</td>
<td>226</td>
<td>226</td>
<td>684</td>
</tr>
<tr>
<td>label_2</td>
<td>684</td>
<td>226</td>
<td>226</td>
</tr>
<tr>
<td>label_3</td>
<td>226</td>
<td>226</td>
<td>684</td>
</tr>
<tr>
<td>label_4</td>
<td>226</td>
<td>226</td>
<td>684</td>
</tr>
<tr>
<td>label_5</td>
<td>684</td>
<td>226</td>
<td>226</td>
</tr>
</tbody>
</table>

#### Coils ID Analysis

<table>
<thead>
<tr>
<th>Coils ID</th>
<th>Category</th>
<th>label_0</th>
<th>label_1</th>
<th>label_2</th>
<th>label_3</th>
<th>label_4</th>
<th>label_5</th>
<th>Length_Hot</th>
<th>Length_Cold</th>
<th>Length_Galv</th>
<th>Cuts_Hot</th>
<th>Cuts_Galv</th>
<th>Thickness_Hot</th>
<th>Thickness_Galv</th>
<th>dynamic_var_00hot</th>
<th>dynamic_var_04hot</th>
<th>dynamic_var_02hot</th>
<th>dynamic_var_03hot</th>
<th>dynamic_var_04hot</th>
</tr>
</thead>
<tbody>
<tr>
<td>00522219</td>
<td>0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>1304601</td>
<td>1304601</td>
<td>1304601</td>
<td>1304601</td>
<td>1304601</td>
<td>1304601</td>
<td>1304601</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Specific technical highlights
Data model
Set of concepts in the ontology
Subset of statistical instances and concepts
Example of instances in ontology
Retrieving models from knowledge data base
Adding models to the knowledge data base
Knowledge usage

1) User selects criteria

2) Ontology engine find best suitable model (using semantic similarity algorithm)

3) Model is copied. Ontology engine maps inputs to inputs relevant to current problem

4) Model can be opened with Rapidminer
Knowledge enrichment

1) User selects a RapidMiner computation process

2) User can add different criteria relevant to its problem

3) Main parameters from the process are automatically extracted

4) New data stored inside ontology