



AI/ML in Practice

Interactivity, Explainability, Scalability
and relationships with Rough Sets

Dominik Ślęzak
December 2020

Interactivity-Explainability-Scalability Solutions

Label in the Loop



CHALLENGES

QUALITY



"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Contributing to the Academic Community



IEEE BigData 2020 Cup: Predicting Escalations in Customer Support

Predicting Escalations in Customer Support is a data mining challenge organized in association with the IEEE BigData 2020 conference. The task is to predict which cases in Information Builders, Inc. (ibi) technical support ticketing system will be escalated in the nearest future by customers. The competition is organized jointly by ibi (<https://www.ibi.com>) and QED Software (<http://www.qed.pl/>).

Manager: Andrzej Janusz (andrzej)

259 teams



2 months, 2 weeks ago



FedCSIS 2020 Challenge: Network Device Workload Prediction

FedCSIS 2020 Data Mining Challenge: Network Device Workload Prediction is the seventh data mining competition organized in association with Conference on Computer Science and Information Systems (<https://fedcis.org/>). This time, the considered task is related to the monitoring of large IT infrastructures and the estimation of their resource allocation. The challenge is sponsored by EMCA Software and Polish Information Processing Society (PTI).

Manager: Andrzej Janusz (andrzej)

162 teams



IEEE BigData 2019 Cup: Suspicious Network Event Recognition

Suspicious Network Event Recognition is a data mining challenge organized in association with IEEE BigData 2019 conference. The task is to decide which alerts should be regarded as suspicious based on information extracted from network traffic logs. The competition is kindly sponsored by Security On-Demand (<https://www.securityondemand.com/>) and QED Software (<http://qed.pl/>).

Manager: Andrzej Janusz (andrzej)

293 teams



Clash Royale Challenge: How to Select Training Decks for Win-rate Prediction

Clash Royale Challenge is the sixth data mining competition organized in association with the Federated Conference on Computer Science and Information Systems (<https://fedcis.org/>). This year, the task is related to the problem of selecting an optimal training data subset for learning how to predict win-rates of the most popular Clash Royale decks. The competition is kindly sponsored by eSensei, QED Software and Polish Information Processing Society (PTI).

Manager: Andrzej Janusz (andrzej)

117 teams



ESENSEI Challenge: Marking Hair Follicles on Microscopic Images

ESENSEI Challenge is a data mining competition, whereby the task is to design an algorithm for accurate marking of follicle positions on microscopic images.

Manager: Andrzej Janusz (andrzej)

49 teams



IJCRRS'15 Data Challenge: Mining Data from Coal Mines

IJCRRS'15 Data Challenge: Mining Data from Coal Mines is a competition organized within a frame of The 2015 International Joint Conference on Rough Sets (IJCRRS'15). It continues the tradition of data mining challenges associated with rough set conferences. This time, the task is related to the problem of monitoring and prediction of dangerous concentrations of methane in longwalls of a Polish coal mine. The competition is sponsored by Research and Development Centre EMAG (<http://www.emag.pl/>) with support from International Rough Set Society.

Manager: Andrzej Janusz (andrzej)

132 teams



PAKDD'15 Data Mining Competition: Gender Prediction Based on E-commerce Data

PAKDD'15 Data Mining Competition: Gender Prediction Based on E-commerce Data is our first competition organized within the frame of The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). We would like to challenge participants with a task of devising effective algorithms for recognizing a gender of e-store clients. A data set for the competition was provided by FPT Group which is also the main sponsor of the awards.

Manager: Andrzej Janusz (andrzej)

414 teams



AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene

AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene is a continuation of the last year's competition organized within the framework of International Symposium on Advances in Artificial Intelligence and Applications (AAIA'15, <https://fedcis.org/aaia>). It is also an integral part of the 2nd Complex Events and Information Modelling workshop (CEIM'15 <https://fedcis.org/ceim>) devoted to the fire protection engineering. This time, the task is related to the problem of recognizing activities carried out by firefighters based on streams of information from body sensor networks. Prizes worth over 4,000 PLN will be awarded to the most successful teams. The contest is sponsored by Polish Information Processing Society (<http://www.pti.org.pl/>), with a support from University of Warsaw (<http://www.mimuw.edu.pl/>) and ICRA project (<http://cra-project.org/>).

Manager: Andrzej Janusz (andrzej)

169 teams



AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service

AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service is organized within the framework of the 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'14, <https://fedcis.org/aaia>), and is an integral part of the 1st Complex Events and Information Modelling workshop (CEIM'14 <https://fedcis.org/ceim>) devoted to the fire protection engineering. The task is related to the problem of extracting useful knowledge from incident reports obtained from the State Fire Service of Poland. Prizes worth over 3,000 USD will be awarded to the most successful teams. The contest is sponsored by Dituel Sp. z o.o. (<http://www.dituel.com.pl/>) and F&K Consulting Engineers (<http://www.fkce.pl/>), with a support from The University of Warsaw (<http://www.mimuw.edu.pl/>) and ICRA project (<http://cra-project.org/>).

Manager: Andrzej Janusz (andrzej)

128 teams



AAIA'18 Data Mining Challenge: Predicting Win-rates of Hearthstone Decks

AAIA'18 Data Mining Challenge is the fifth competition organized within the framework of International Symposium Advances in Artificial Intelligence and Applications (<https://fedcis.org/2018/aaia>). This time, the task is to assess win-rates of Hearthstone decks in games played between AI bots. The competition is kindly sponsored by Silver Bullet Labs, eSensei and Polish Information Processing Society (PTI).

Manager: Andrzej Janusz (andrzej)

217 teams



2 years, 7 months ago



AAIA'17 Data Mining Challenge: Helping AI to Play Hearthstone

AAIA'17 Data Mining Challenge is the fourth data mining competition organized within the framework of International Symposium Advances in Artificial Intelligence and Applications (<https://fedcis.org/2017/aaia>). This time, the task is to come up with an efficient prediction model which would help AI to play the game of Hearthstone: Heroes of Warcraft. The competition is kindly sponsored by Silver Bullet Solutions and Polish Information Processing Society (PTI).

Manager: Andrzej Janusz (andrzej)

369 teams



3 years, 7 months ago



ISMIS'17 Data Mining Competition: Trading Based on Recommendations

ISMIS 2017 Data Mining Competition is a challenge organized using the KnowledgePit platform at the 23rd International Symposium on Methodologies and Intelligent Systems, held at Warsaw University of Technology, Poland, on June 26-29, 2017. The task is to come up with a strategy for investing in a stock market based on recommendations provided by different experts. The competition is kindly sponsored by mBank S.A. and Tipranks, with a support from ISMIS 2017 organizers.

Manager: Andrzej Janusz (andrzej)

177 teams



3 years, 10 months ago



AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines

AAIA'16 Data Mining Challenge is the third data mining competition associated with International Symposium on Advances in Artificial Intelligence and Applications (AAIA'16, <https://fedcis.org/2016/aaia>) which is a part of FedCSIS conference series. This time, the task is related to the problem of predicting periods of increased seismic activity which may cause life-threatening accidents in underground coal mines. Prizes worth over 3,000 USD will be awarded to the most successful teams. The contest is sponsored by Research and Development Centre EMAG (<http://emag.pl>) with support from Polish Information Processing Society (<http://www.pti.org.pl>) and Dituel Sp. z o.o. (<http://www.dituel.pl/>).

Manager: Andrzej Janusz (andrzej)

251 teams



4 years, 9 months ago




Scientific Research and Industry Collaborations


Predicting Escalations in Customer Support:
Analysis of Data Mining Challenge Results

Dominik Ślęzak, Andrzej Janusz, Daniel Kałuża




 Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 105–114 ISSN 2300-5963 ACSIS, Vol. 21
DOI: 10.15439/2020F189

Integrated Human Tracking Based on Video and Smartphone Signal Processing within the Arahub System

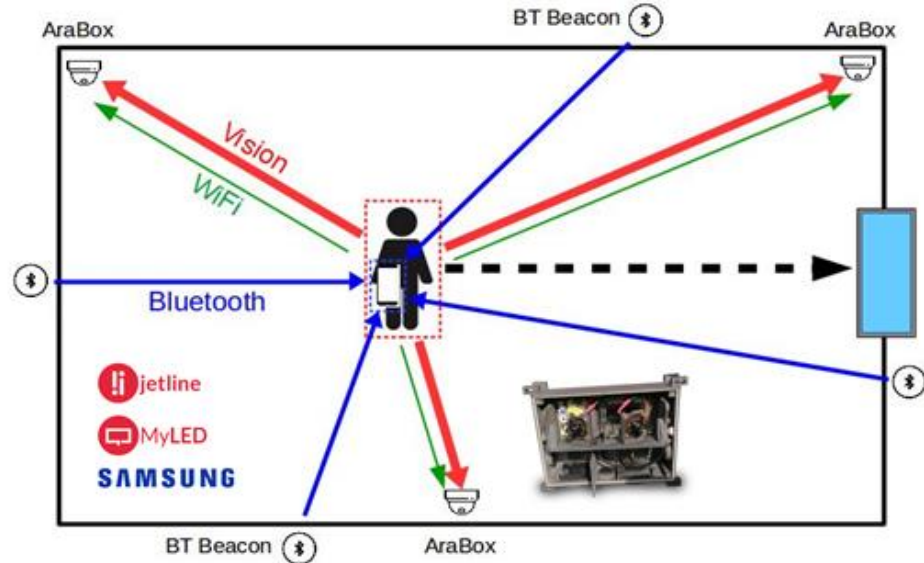
Jan Ludziejewski, Lukasz Grad
Uniwersytet Warszawski
Email: [jan.ludziejewski, lukasz.grad]@mimuw.edu.pl

Lukasz Przebinda
Arahub & Myled
Email: l.przebinda@myled.pl


Tomasz Tajmajer
QED Software
Email: tomasz.tajmajer@qed.pl

Abstract—Embedded platforms with GPU acceleration, designed for performing machine learning on the edge, enabled the creation of inexpensive and pervasive computer vision systems. Smartphones are nowadays widely used for profiling and tracking in marketing, based on WiFi data or beacon-based positioning systems. We present the Arahub system, which ...



In this paper the overall architecture of the Arahub system is described. We provide insights into particular elements of the system and methods used. We also present preliminary results, which we were able to obtain in real-life environments.



The diagram illustrates the Arahub system architecture. A central user (represented by a person icon) is surrounded by several components: two AraBox units (top-left and top-right), two BT Beacons (top and bottom), and a smartphone (right). The AraBox units are connected to the user via Vision (red arrows) and WiFi (green arrows). The BT Beacons are connected to the user via Bluetooth (blue arrows). The smartphone is connected to the user via Bluetooth (blue arrows). The AraBox units are also connected to the BT Beacons. The diagram also shows a Samsung smartphone and a MyLED AraBox unit.


Reinventing Infobright's Concept of
Rough Calculations on Granulated Tables
for the Purpose of Accelerating
Modern Data Processing Frameworks

Mateusz Wniak¹, Sebastian Stawicki¹, Dominik Ślęzak¹
¹Institute of Informatics, University of Warsaw, Poland
QED Software, Poland



Theory of Rough Sets

- Rough Set Approximations
- Simple data and information representations
- Approximating complex phenomena with compacted and intuitive models

Publications authored by Professor Pawlak have still the highest number of citations when considering all scientists affiliated in Poland (<https://data.mendeley.com/datasets/btchxktyw/2>)



Chapter 1

Professor Zdzisław Pawlak (1926-2006): Founder of the Polish School of Artificial Intelligence

Andrzej Skowron*, Mihir Kr. Chakraborty, Jerzy Grzymała-Busse, Victor Marek, Sankar K. Pal, James F. Peters, Grzegorz Rozenberg, Dominik Ślęzak, Roman Słowiński, Shusaku Tsumoto, Alicja Wakulicz-Deja, Guoyin Wang, and Wojciech Ziarko

*He was not just a great scientist – he was also
a great human being.*

– Lotfi A. Zadeh, April 2006

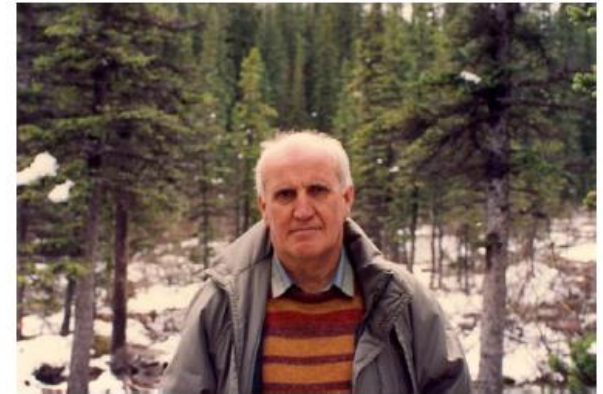
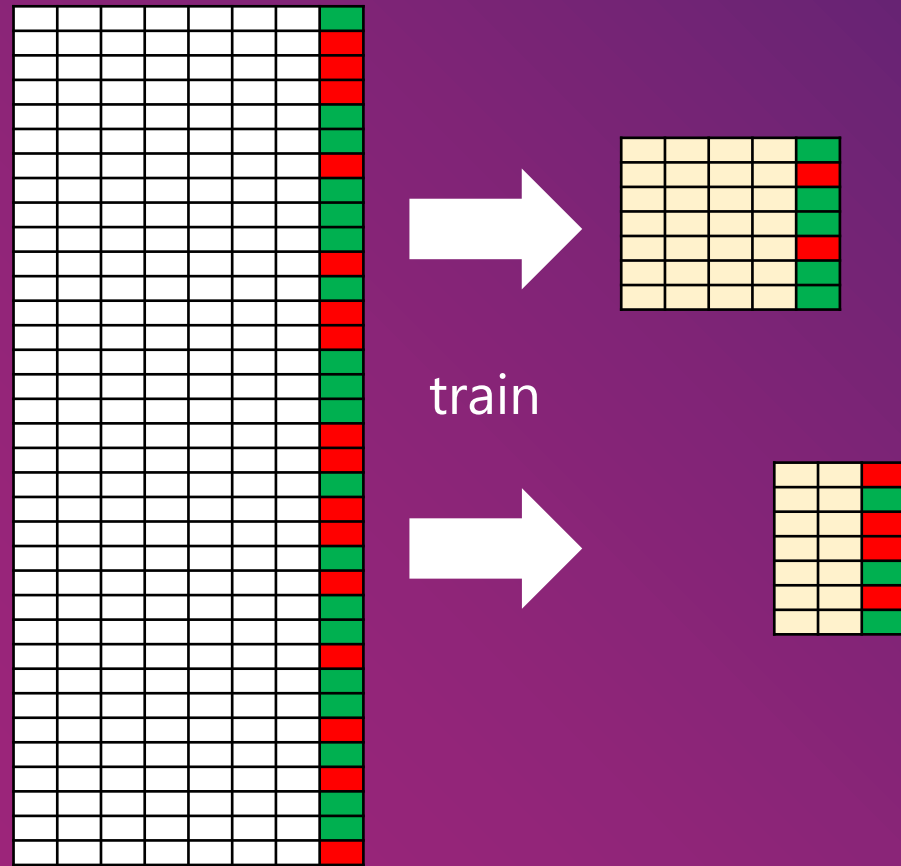


Fig. 1.1. Zdzisław Pawlak

Andrzej Skowron
Institute of Mathematics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland
e-mail: skowron@mimuw.edu.pl

* Corresponding author.

Ensembles of Reducts



Information Sciences 451–452 (2018) 112–133

Contents lists available at ScienceDirect

Information Sciences

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

ELSEVIER

Check for updates

A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines

Dominik Ślęzak^{a,*}, Marek Grzegorowski^a, Andrzej Janusz^a, Michał Kozielski^b, Sinh Hoa Nguyen^c, Marek Sikora^{b,d}, Sebastian Stawicki^a, Łukasz Wróbel^{b,d}

^aInstitute of Informatics, University of Warsaw, ul. Banacha 2, Warsaw 02-097, Poland
^bInstitute of Informatics, Silesian University of Technology, ul. Akademicka 16, Gliwice 44-100, Poland
^cPolish-Japanese Academy of Information Technology, ul. Koszykowa 86, Warsaw 02-008, Poland
^dInstitute of Innovative Technologies EMAC, ul. Leopolda 31, Katowice 40-189, Poland

ARTICLE INFO

Article history:
Received 31 May 2017
Revised 18 February 2018
Accepted 3 April 2018
Available online 4 April 2018

Keywords:
Sensor data processing
Methane concentration forecasting
Sliding-window feature engineering
Feature subset ensemble selection

ABSTRACT

We introduce a new approach for learning forecasting models over large multi-sensor data sets, including the steps of sliding-window-based feature extraction and rough-set-inspired feature subset ensemble selection. We show how to integrate this approach with the major data-processing-related components of DISESOR – a decision support system which is a coherent and complete framework for exploring streams of sensor readings registered in underground coal mines. As a case study, we report our experiments related to the task of methane concentration forecasting. The contributions in this paper refer to both the analysis how the nature of sensor readings influenced the architecture of the developed system and the empirical proof that the designed methods for data processing and analytics turned out to be efficient in practice.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

In the last decade, intensive growth in capabilities and popularity of analytical environments containing data mining solutions has been observed. Marketing, insurance, banking and finance, trade (especially e-commerce) and health care are the most popular applications. Less frequently, data mining methods are used to analyze and supervise industrial processes. The industrial monitoring systems usually produce multivariate streams of sensor readings for which performing standard preprocessing steps (such as data integration, data cleaning, feature extraction, etc.) is quite challenging. It is also difficult to construct and maintain forecasting models that should be used in an on-line fashion in industrial decision support systems. Nevertheless, potential benefits coming from intelligent utilization of this data source are truly huge.

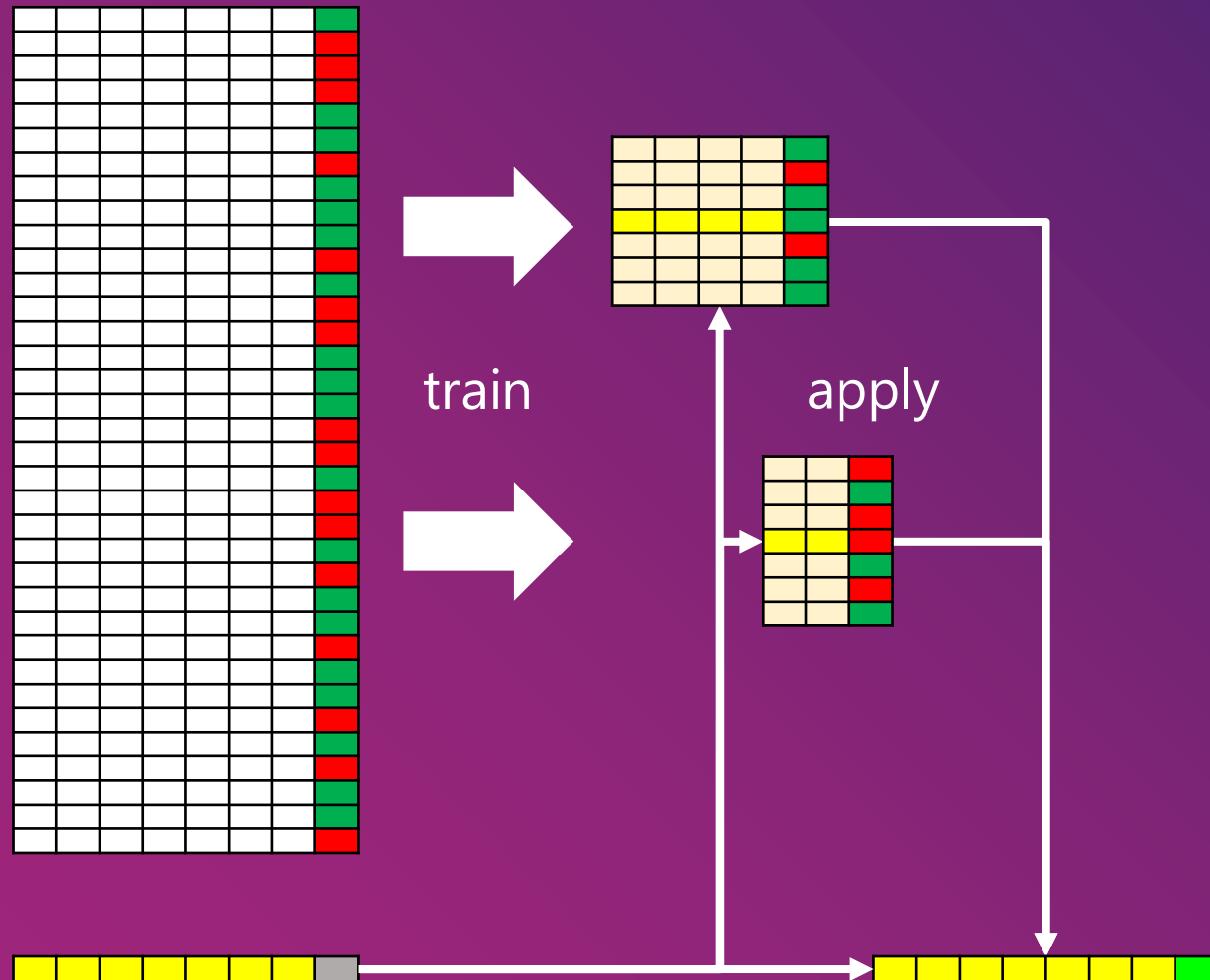
In this paper, we investigate new approaches for learning forecasting models from multi-sensor data, for the purposes of monitoring natural hazards and industrial processes. We discuss them from the viewpoint of our decision support system – called DISESOR – which comprises of the expert system shell with the knowledge base that can be used together with the data incoming on-line, the feature engineering module that can derive the most meaningful statistics describing multivariate

* Corresponding author.
E-mail address: slęzak@mimuw.edu.pl (D. Ślęzak).

<https://doi.org/10.1016/j.ins.2018.04.026>
0020-0255/© 2018 Elsevier Inc. All rights reserved.



Ensembles of Reducts



A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines

Dominik Ślęzak^{a,*}, Marek Grzegorowski^a, Andrzej Janusz^a, Michał Kozielski^b, Sinh Hoa Nguyen^c, Marek Sikora^{b,d}, Sebastian Stawicki^a, Łukasz Wróbel^{b,d}

^aInstitute of Informatics, University of Warsaw, ul. Banacha 2, Warsaw 02-097, Poland

^bInstitute of Informatics, Silesian University of Technology, ul. Akademicka 16, Gliwice 44-100, Poland

^cPolish-Japanese Academy of Information Technology, ul. Koszykowa 86, Warsaw 02-008, Poland

^dInstitute of Innovative Technologies EMAG, ul. Leopolda 31, Katowice 40-189, Poland

ARTICLE INFO

Article history:

Received 31 May 2017

Revised 18 February 2018

Accepted 3 April 2018

Available online 4 April 2018

Keywords:

Sensor data processing

Methane concentration forecasting

Sliding-window feature engineering

Feature subset ensemble selection

ABSTRACT

We introduce a new approach for learning forecasting models over large multi-sensor data sets, including the steps of sliding-window-based feature extraction and rough-set-inspired feature subset ensemble selection. We show how to integrate this approach with the major data-processing-related components of DISESOR – a decision support system which is a coherent and complete framework for exploring streams of sensor readings registered in underground coal mines. As a case study, we report our experiments related to the task of methane concentration forecasting. The contributions in this paper refer to both the analysis how the nature of sensor readings influenced the architecture of the developed system and the empirical proof that the designed methods for data processing and analytics turned out to be efficient in practice.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

In the last decade, intensive growth in capabilities and popularity of analytical environments containing data mining solutions has been observed. Marketing, insurance, banking and finance, trade (especially e-commerce) and health care are the most popular applications. Less frequently, data mining methods are used to analyze and supervise industrial processes. The industrial monitoring systems usually produce multivariate streams of sensor readings for which performing standard preprocessing steps (such as data integration, data cleaning, feature extraction, etc.) is quite challenging. It is also difficult to construct and maintain forecasting models that should be used in an on-line fashion in industrial decision support systems. Nevertheless, potential benefits coming from intelligent utilization of this data source are truly huge.

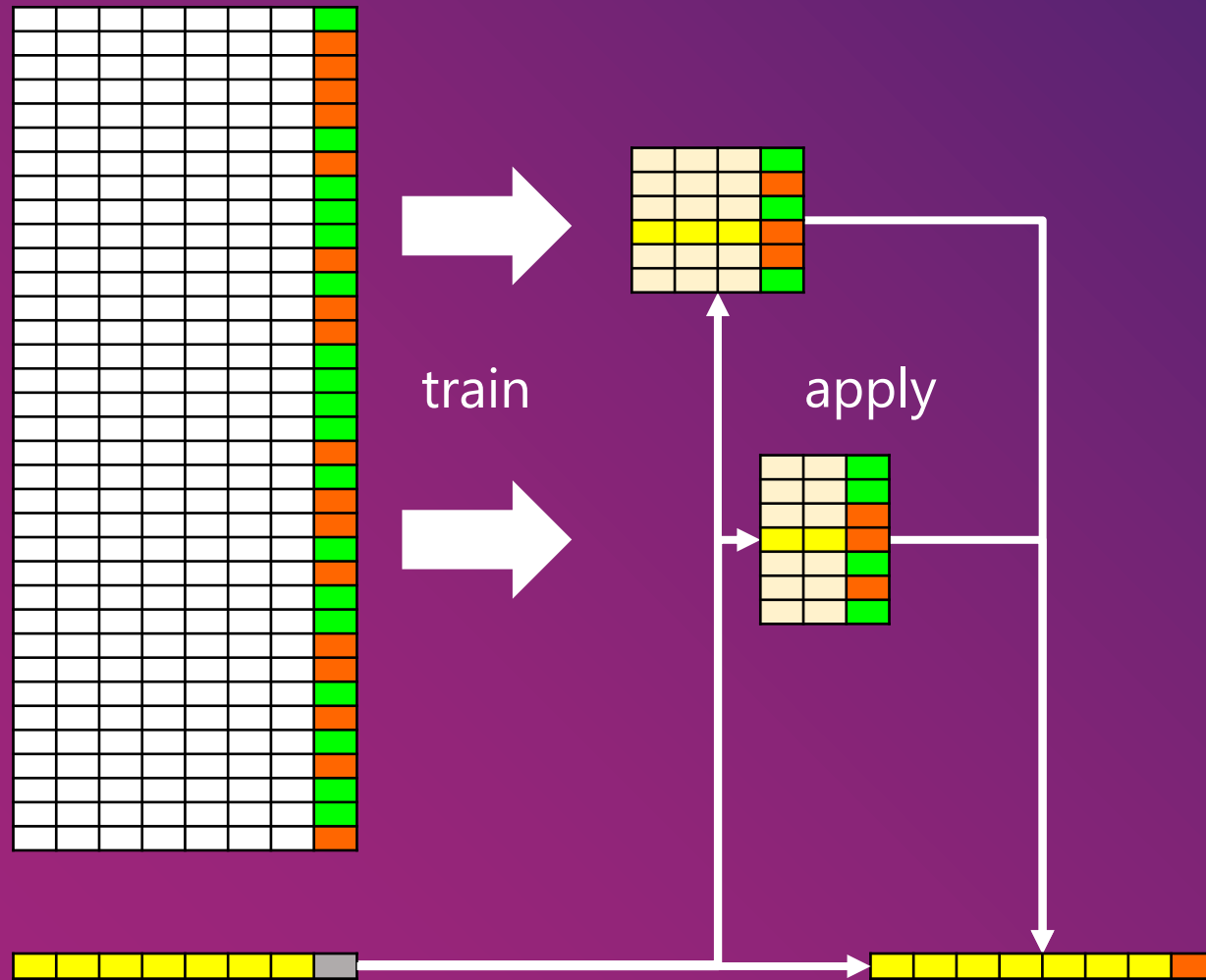
In this paper, we investigate new approaches for learning forecasting models from multi-sensor data, for the purposes of monitoring natural hazards and industrial processes. We discuss them from the viewpoint of our decision support system – called DISESOR – which comprises of the expert system shell with the knowledge base that can be used together with the data incoming on-line, the feature engineering module that can derive the most meaningful statistics describing multivariate

* Corresponding author.

E-mail address: slszak@mimuw.edu.pl (D. Ślęzak).



Reduct Approximators



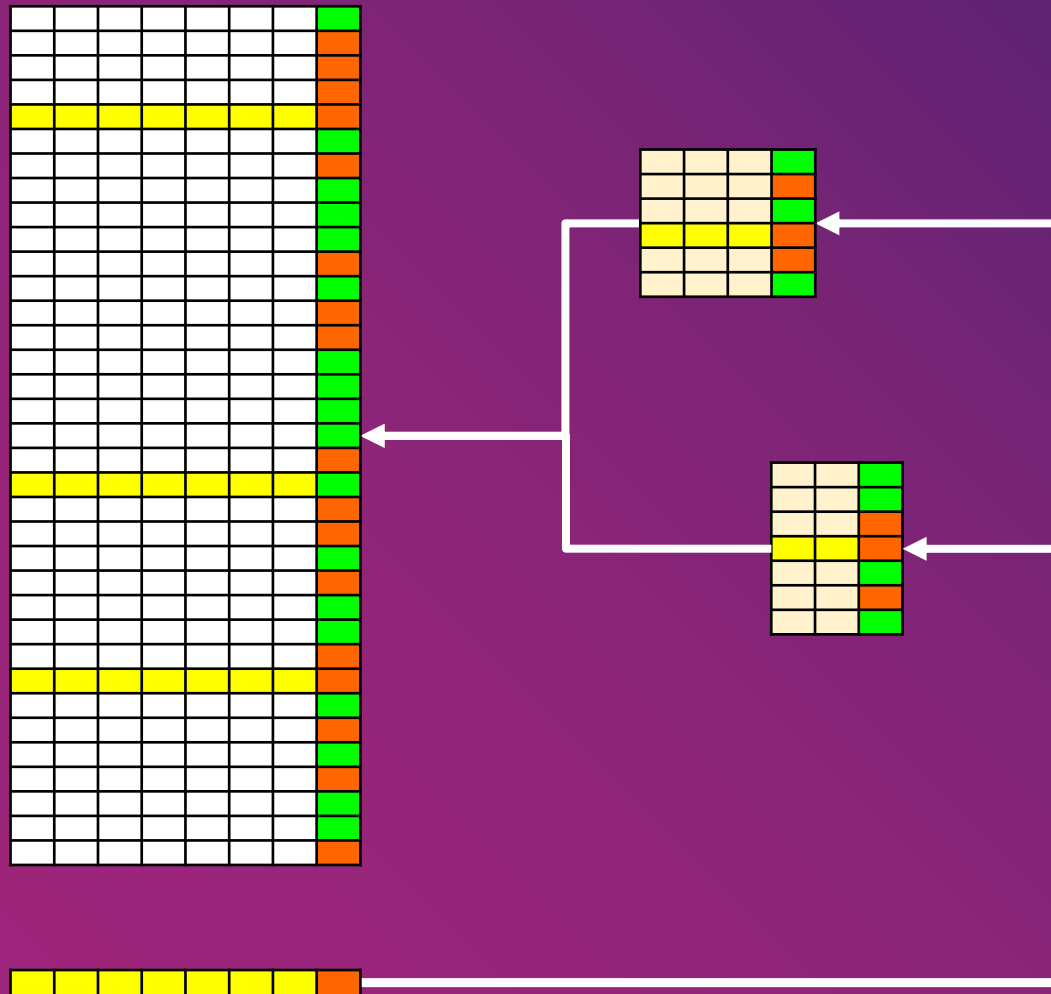
EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models



Reduct Approximators



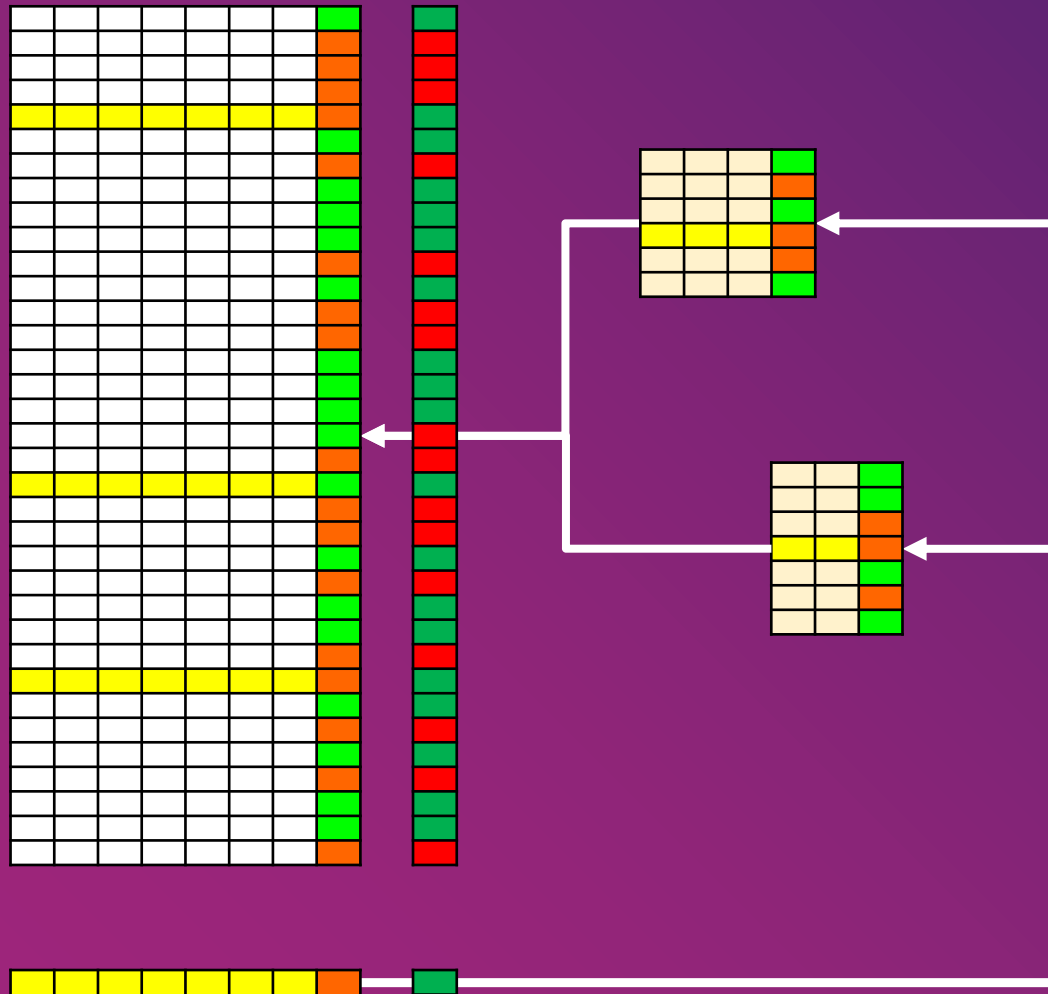
EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models



BrightBox Diagnostics



IEEE BigData 2020 Cup: Predicting Escalations in Customer Support

Predicting Escalations in Customer Support is a data mining challenge organized in association with the IEEE BigData 2020 conference. The task is to predict which cases in Information Builders, Inc. (ibi) technical support ticketing system will be escalated in the nearest future by customers. The competition is organized jointly by ibi (<https://www.ibi.com>) and QED Software (<http://www.qed.pl/>).



2 months, 2 weeks ago

Manager: Andrzej Janusz (andrzej)

259 teams



FedCSIS 2020 Challenge: Network Device Workload Prediction

FedCSIS 2020 Data Mining Challenge: Network Device Workload Prediction is the seventh data mining competition organized in association with Conference on Computer Science and Information Systems (<https://fedcsis.org/>). This time, the considered task is related to the monitoring of large IT infrastructures and the estimation of their resource allocation. The challenge is sponsored by EMCA Software and Polish Information Processing Society (PTI).



6 months, 1 week ago

Manager: Andrzej Janusz (andrzej)

162 teams



IEEE BigData 2019 Cup: Suspicious Network Event Recognition

Suspicious Network Event Recognition is a data mining challenge organized in association with IEEE BigData 2019 conference. The task is to decide which alerts should be regarded as suspicious based on information extracted from network traffic logs. The competition is kindly sponsored by Security On-Demand (<https://www.securityondemand.com/>) and QED Software (<http://qed.pl/>).



1 year, 2 months ago

Manager: Andrzej Janusz (andrzej)

293 teams



Clash Royale Challenge: How to Select Training Decks for Win-rate Prediction

Clash Royale Challenge is the sixth data mining competition organized in association with the Federated Conference on Computer Science and Information Systems (<https://fedcsis.org/>). This year, the task is related to the problem of selecting an optimal training data subset for learning how to predict win-rates of the most popular Clash Royale decks. The competition is kindly sponsored by eSensei, QED Software and Polish Information Processing Society (PTI).



1 year, 6 months ago

Manager: Andrzej Janusz (andrzej)

117 teams



ESENSEI Challenge: Marking Hair Follicles on Microscopic Images

ESENSEI Challenge is a data mining competition, whereby the task is to design an algorithm for accurate marking of follicle positions on microscopic images.



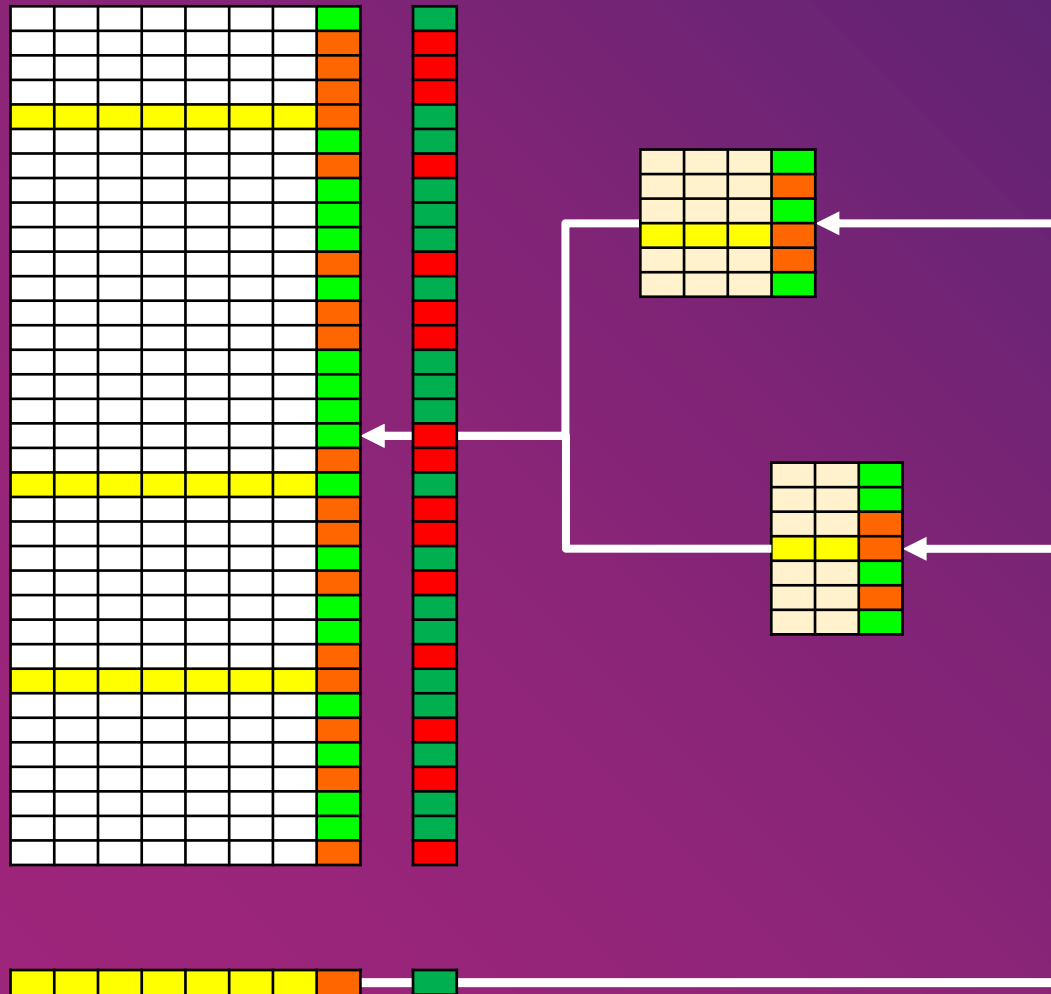
2 years, 6 months ago

Manager: Andrzej Janusz (andrzej)

49 teams



Computational Challenges



- Diagnose a big volume of new cases at a time
- Extract neighborhoods of every new case from a very big volume of historical cases
- But first of all, learn a good ensemble of reducts from that very big volume
- And finally, make it all a repeatable, adaptive and production-ready process



Design of InfoFrames

InfoData

- compressed data
 - row groups
- efficient data access
- multi-threaded data processing

InfoIndex

- basics stats
- navigation function
- support efficient analysis (e.g. filtering)

InfoInsight

- granulated data summaries
- reducing the size of the data
- preserving useful information
- compressed representations of the original data



SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Design of InfoFrames

InfoData

- compressed data
 - row groups
- efficient data access
- multi-threaded data processing

InfoIndex

- basics stats
- navigation function
- support efficient analysis (e.g. filtering)

InfoInsight

- granulated data summaries
- reducing the size of the data
- preserving useful information
- compressed representations of the original data



SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows

Parquet

Initial release	13 March 2013; 7 years ago
Stable release	2.8.0 / 13 January 2020; 11 months ago ^[1]
Repository	git-wip-us.apache.org/repos/asf/parquet-mr.git
Written in	Java (reference implementation) ^[2]
Operating system	Cross-platform
Type	Column-oriented DBMS
License	Apache License 2.0
Website	parquet.apache.org



Design of InfoFrames

InfoData

- compressed data
 - row groups
- efficient data access
- multi-threaded data processing

InfoIndex

- basics stats
- navigation function
- support efficient analysis (e.g. filtering)

InfoInsight

- granulated data summaries
- reducing the size of the data
- preserving useful information
- compressed representations of the original data



SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Initial release 13 March 2013; 7 years ago

Stable release 2.8.0 / 13 January 2020; 11 months ago^[1]

Repository [git-wip-us.apache.org/repos/asf/parquet-mr.git](https://github.com/apache/parquet-mr)

Written in Java (reference implementation)^[2]

Operating system Cross-platform

Type Column-oriented DBMS

License Apache License 2.0

Website parquet.apache.org

2020-02-10, last modified: 2020-02-10 by Wojciech Ojeda

Reinventing Infobright's Concept of Rough Calculations on Granulated Tables for the Purpose of Accelerating Modern Data Processing Frameworks

Marcin Wasił¹, Sebastian Szustek¹, Dominik Szustek¹
¹Institute of Informatics, University of Warsaw, Poland
²IGIS Informatics, Poland

Abstract. We present an approach to data and information management based on the Infobright Community Edition (ICE) multi-table data processing framework. Apache Parquet and HDFS Block of Data Objects (BDO) are used to store the data of existing queries based on rough calculations on granulated data. We analyze the way of the data organization in ICE to maximize the performance of such kind of operations and compare the performance of Parquet and HDFS. We also analyze the performance of Parquet and HDFS. We also analyze the performance of Parquet and HDFS.

1. INTRODUCTION

Infobright¹ was developed as an analytical database engine representing the idea of data granulation and rough sets in order to minimize the need of accuracy and discretization of the data while working SQL queries [1], [2]. The approach is reported in this paper as based on an open source extension of the data processing framework called Infobright Community Edition (ICE) [3].

ICE stores the data as a form of collection of granulated data objects (GDOs). Each GDO is a collection of data objects (DOs) which are organized in a form of a data pack (DP). Each DP is a collection of data objects (DOs) which are organized in a form of a data pack (DP). Each DP is a collection of data objects (DOs) which are organized in a form of a data pack (DP).

The goal of this paper is to show that Infobright's approach can be used to accelerate data operations in other computing frameworks. For this purpose, we compare with an abstract notion of a granulated table, whereby rows correspond to rows and columns correspond to column types of statistics calculated for these rows. We compare the performance of the corresponding sets of query operations in two selected environments: Apache Parquet² and HDFS³. Both of them provide a library supporting APIs to perform complex data operations on the distributed files of information granulation (e.g. granulation, whereby granules (atoms) are organized corresponding to the packs and data packs as well as information granulation corresponding to their statistics and

¹Infobright Group (https://www.infobright.com/)
²Apache Parquet (https://parquet.apache.org/)
³Hadoop Distributed File System (https://hadoop.apache.org/docs/en/3.1.0/hdfs.html)

PLA-150-451-13163-10-2020-02-10 1401



Design of InfoFrames

InfoData

- compressed data
 - row groups
- efficient data access
- multi-threaded data processing

InfoIndex

- basics stats
- navigation function
- support efficient analysis (e.g. filtering)

InfoInsight

- granulated data summaries
- reducing the size of the data
- preserving useful information
- compressed representations of the original data



SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Initial release 13 March 2013; 7 years ago

Stable release 2.8.0 / 13 January 2020; 11 months ago^[1]

Repository [git-wip-us.apache.org/repos/asf/parquet-mr](https://github.com/apache/parquet-mr)^[git]

Written in Java (reference implementation)^[2]

Operating system Cross-platform

Type Column-oriented DBMS

License Apache License 2.0

Website parquet.apache.org^[g]

2022-03-03, last modified by Big Data Blog

Reinventing Inforight's Concept of Rough Calculations on Granulated Tables for the Purpose of Accelerating Modern Data Processing Frameworks

Marek Wasik¹, Sebastian Szewczyk¹, Dominik Szpak^{1,2}
¹Institute of Informatics, University of Warsaw, Poland
²AGH University, Poland

Abstract. We present an approach to data and information processing based on the Inforight's Concept of Rough Calculations on Granulated Tables (ICRT). This approach is based on the idea of rough sets and rough approximations. The idea of rough sets is based on the idea of granulation. The idea of rough approximations is based on the idea of granulation. The idea of rough sets is based on the idea of granulation. The idea of rough approximations is based on the idea of granulation.

1. INTRODUCTION

Inforight's Concept of Rough Calculations on Granulated Tables (ICRT) is a novel approach to data and information processing based on the idea of rough sets and rough approximations. The idea of rough sets is based on the idea of granulation. The idea of rough approximations is based on the idea of granulation.

2. RELATED WORK

The idea of rough sets is based on the idea of granulation. The idea of rough approximations is based on the idea of granulation.

3. CONCLUSION

The idea of rough sets is based on the idea of granulation. The idea of rough approximations is based on the idea of granulation.

Security On-Demand is the only MSSP/MDR company in the world that utilizes AQ Technology.

WHAT IS AQ TECHNOLOGY?

SOD's ground-breaking AQ Technology™ is based on Rough Set mathematics and Differentiated Intelligence models. AQ Technology™ is a data reduction engine that filters the data and the application which provides analysis of indicators 10x-faster than querying the database directly.

OUR PATENTED TECHNOLOGY

Security On-Demand is the only MSSP/MDR company in the world that utilizes AQ Technology™. With our patented SOD we protect the user for every computer on the managed security services industry.

THREAT DETECTION ON A MASSIVE SCALE

View our latest threat intelligence reports and see how AQ Technology™ adds the edge.

A NEW REVOLUTIONARY APPROACH TO THREAT ANALYTICS

Patented AQ Technology™ powers all intelligence in advanced threat detection.

RAPID ANALYSIS OF THREAT DATA

See how AQ Technology™ processes 100,000+ threat events in 100ms from data and a 100+ TB of data.



Design of InfoFrames

InfoData

- compressed data
 - row groups
- efficient data access
- multi-threaded data processing

InfoIndex

- basics stats
- navigation function
- support efficient analysis (e.g. filtering)

InfoInsight

- granulated data summaries
- reducing the size of the data
- preserving useful information
- compressed representations of the original data



SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



Initial release 13 March 2013; 7 years ago

Stable release 2.8.0 / 13 January 2020; 11 months ago^[1]

Repository [git-wip-us.apache.org/repos/asf/parquet-mr](https://github.com/apache/parquet-mr)^[git]

Written in Java (reference implementation)^[2]

Operating system Cross-platform

Type Column-oriented DBMS

License Apache License 2.0

Website parquet.apache.org^[a]

2022-03-02, last modified by Big Data Blog Team

Reinventing Inforight's Concept of Rough Calculations on Granulated Tables for the Purpose of Accelerating Modern Data Processing Frameworks

Marek Wasił¹, Sebastian Szewczyk¹, Dominik Szewczyk¹
¹Institute of Informatics, University of Warsaw, Poland
²IGIS, Informatics, Poland

Abstract. We present an approach to data and information processing based on the Inforight's Concept of Rough Calculations on Granulated Tables (ICGT). This approach is based on the idea of rough sets and rough approximations. It is designed to accelerate modern data processing frameworks by reducing the size of the data and preserving useful information. The approach is based on the idea of rough sets and rough approximations. It is designed to accelerate modern data processing frameworks by reducing the size of the data and preserving useful information. The approach is based on the idea of rough sets and rough approximations. It is designed to accelerate modern data processing frameworks by reducing the size of the data and preserving useful information.

Security On-Demand is the only MSSP/MDR company in the world that utilizes AQ Technology.

WHAT IS AQ TECHNOLOGY?

Security On-Demand (SOD) is based on Inforight's ground-breaking AQ Technology. It is based on Inforight's ground-breaking AQ Technology. It is based on Inforight's ground-breaking AQ Technology. It is based on Inforight's ground-breaking AQ Technology.

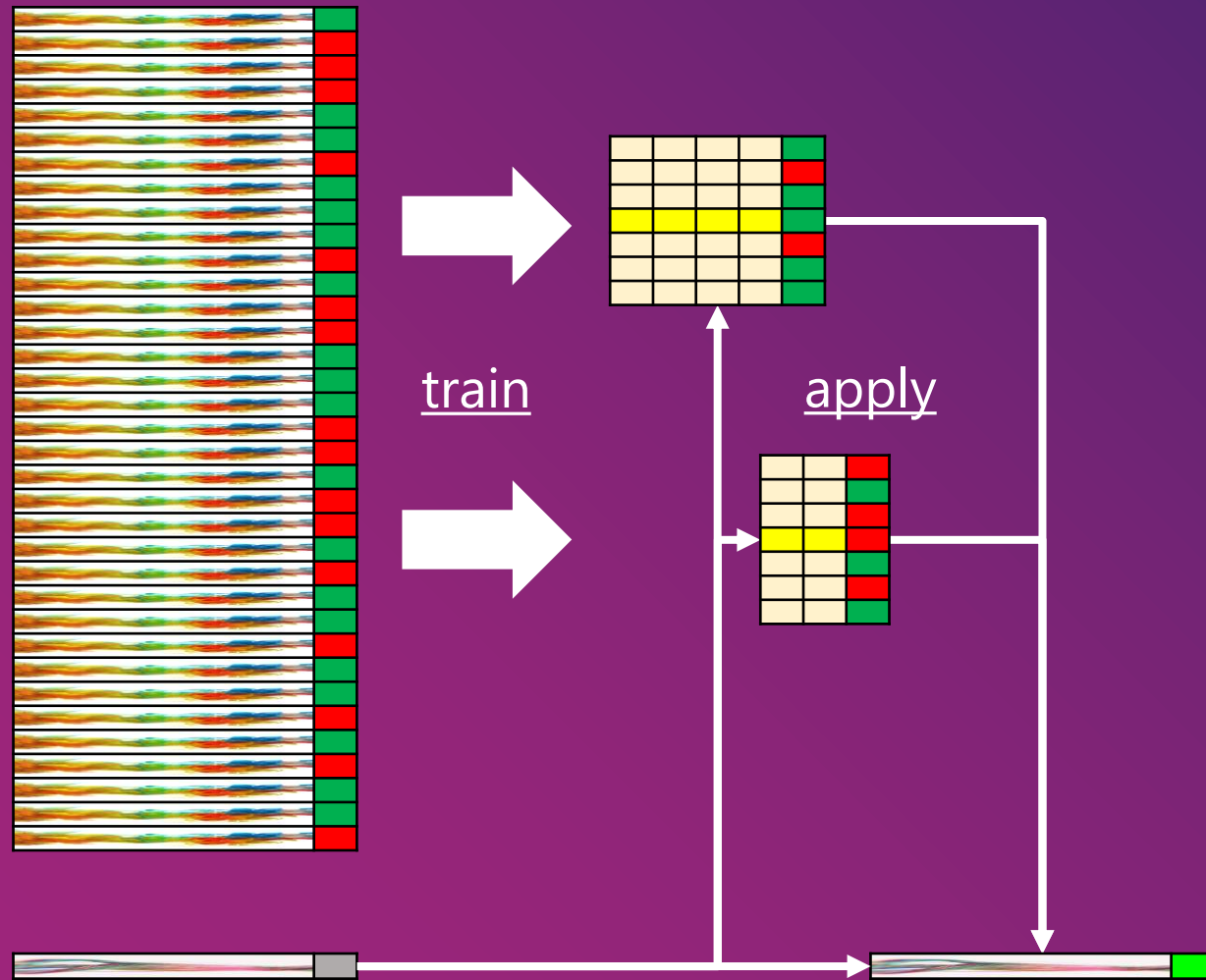
Small Summaries for Big Data

Graham Cormode • Ke Yi



InfoInsight can be filled with parameters of Gaussian mixture „hats“ approximating single dimensions + hidden variable S such that $\sum_{i=1, \dots, n} \text{Ent}(G_i | S) - \text{Ent}(G_1 \dots G_n | S) \rightarrow \min; \text{Ent}(s) \rightarrow \min$

Feature Magazine



- At its growth phase, the reduct creation algorithm repeats the following steps:
 1. Take a random subset of features / attributes
 2. Evaluate its elements
 3. Add the best element to a reduct candidate that is being constructed
- Step 1 can draw random subsets from a space of possible features, instead of a closed set of columns



Feature Magazine

- The space of features needs to be prescribed prior to the beginning of the reduct construction process
- During the process, we may try to get the users involved into the feature selection decision-making
- We may involve the users also at a stage of creating feature candidates, not only selecting them

Label in the Loop



QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling



Interacting with Data



Label in the Loop



QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling



Interacting with Data

- (Explicit) creation of new cases by humans
- (Implicit) creation of features by humans
- AI/ML needs both – cases and features



Label in the Loop



QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling



AI ≠ ML (referring to Video Games as an illustration)

All Games > Strategy Games > Tactical Troops: Anthracite Shift

Tactical Troops: Anthracite Shift

Community Hub



Tactical Troops: Anthracite Shift is an indie turn-based tactical science-fiction game set on the beautiful yet dangerous planet of Anthracite.



The image displays a promotional page for the game 'Tactical Troops: Anthracite Shift'. It features a large top-down tactical screenshot on the left, a game cover on the right, and a descriptive paragraph. Below the cover is a diagram with three overlapping triangles and soldier icons. At the bottom left, there is a video player with a play button and a thumbnail of a game scene.



AI ≠ ML (referring to Video Games as an illustration)

All Games > Strategy Games > Tactical Troops: Anthracite Shift

Tactical Troops: Anthracite Shift

Community Hub





Tactical Troops: Anthracite Shift is an indie turn-based tactical science-fiction game set on the beautiful yet dangerous planet of Anthracite.




At the bottom of the game page, there is a row of five small thumbnail images showing different in-game scenes.

Label in the Loop



CHALLENGES

QUALITY 

"No more garbage data" - models are only as good as the provided data


Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY 

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows



AI ≠ ML (referring to Video Games as an illustration)

All Games > Strategy Games > Tactical Troops: Anthracite Shift

Tactical Troops: Anthracite Shift

Community Hub



Tactical Troops: Anthracite Shift is an indie turn-based tactical science-fiction game set on the beautiful yet dangerous planet of Anthracite.




Introducing LogDL – Log Description Language for Insights from Complex Data


Maciej Świechowski
QED Software, Warsaw, Poland
Email: maciej.swiechowski@qed.pl


Dominik Ślęzak
Institute of Informatics
University of Warsaw, Poland


Abstract—We propose a new logic-based language called *Log Description Language (LogDL)*, designed to be a medium for the knowledge discovery workflows over complex data sets. It makes it possible to operate with the original data along with machine-learning-driven insights expressed as facts and rules, regarded as so-called *descriptive logs* characterizing the observed processes in real or virtual environments. LogDL is inspired by the research at the border of AI and games, precisely by *Game Description Language (GDL)* that was developed for *General Game Playing (GGP)*. We emphasize that such formal frameworks for analyzing the gameplay data are a good prerequisite for the case of real, “not digital” processes. We also refer to *Fogs of War (FoW)* – our upcoming project related to AI in video games with limited information – whereby LogDL will be used as well.

Though the original data shall remain unstructured or multi-structured, the layer of insights can take a form of collections of well-established facts, rules and formulas expressed in LogDL – the aforementioned d-logs describing the observed processes and activities in real or virtual environments. Figure 1 illustrates the usage of LogDL and how it can be involved in various data-related operations and activities.

LogDL provides building blocks (e.g. facts, operators, rules) which algorithms may use, constraints (expressed by means of e.g. domains and rules) within which they operate and some built-in concepts such as time that can be interpreted automatically...







QUALITY

"No more garbage data" - models are only as good as the provided data

Maintaining model performance through time

Better / faster / cheaper data labelling

EXPLAINABILITY

Explainable models increase the quality of decision-making

Understand how data quality affects the prediction models

SCALABILITY

Implementing AI/ML where such a possibility was throttled by processing speed requirements or data scale

Enabling machine learning scalability for big data and/or big data flows

CHALLENGES

- New project arrival (Q2 2021)

FOGS OF WAR

- AI players dealing with incomplete information
- A need of visualizing AI reasoning to humans
- „ML Turing Test” for assessing AI’s humanity



Thank you!

Dominik Ślęzak
December 2020