DETECTING ANOMALIES OVER MESSAGE STREAMS IN RAILWAY COMMUNICATION SYSTEMS

Lucas Foulon\textsuperscript{1,3}, Serge Fenet\textsuperscript{1}, Christophe Rigotti\textsuperscript{2}, Denis Jouvin\textsuperscript{3}

\textsuperscript{1}Université Claude Bernard Lyon 1, CNRS, LIRIS, UMR5205
\textsuperscript{2}Université Lyon, INSA Lyon, CNRS, INRIA, LIRIS, UMR5205
\textsuperscript{3}Production Ferroviaire, SNCF Mobilité, DSI Voyageurs

GOALS

• Monitor on real-time the proper functioning of the information system
• Support high volume of streaming data
• Warn when an anomaly occurs

OUR DATA

• Traces containing information about messages flowing in the information system: number of messages, latency between different checkpoints, ...
• Built by analyzing the content of the data stream: Sent/Received timestamp, type of device/service, ...
• Interfaced with the central platform of the SNCF IS (CanalTrain) through ELK open source products

RESULTS

• Reduced complexity allowing the efficient use of the CFOF score on high volume data streams
• High quality of the estimated score
• Real-time detection of IS anomalies
• One parameter controlling the detection
• Incremental update of the tree

IN PROGRESS

• From tree to forest to reduce dimensions and accelerate the computing
• Multi-scale and multi-indicators anomaly detection
• Testing the robustness to regime changes

METHOD

Use of CFOF anomaly measure [Angiulli, ECML PKDD 2017]
• Unsupervised
• Based on the structure of the local neighborhood
• Adapted to high dimension data
• But not adapted to data streams,

Use of the iSAX indexation tree [Shieh & Keogh, DAMI 2009]
• Based on a modification of the SAX discretization
• Suited for times series indexation and similarity search
• Efficient access using distance boundings
• Support Dynamic Time Warping, weighting, and very high volumes (billion time series)

Proposition: 

• exploit the properties of the iSAX tree to accelerate the computing of the CFOF score in order to apply it to voluminous data streams

• Reduced complexity allowing the efficient use of the CFOF score on high volume data streams
• High quality of the estimated score
• Real time detection of IS anomalies
• One parameter controlling the detection
• Incremental update of the tree

4th AALTD@ECML/PKDD 2019 – 20/09/2019