

Overview delle linee tematiche principali

Raggruppamento nelle principali "linee tematiche" su cui studentesse e studenti negli anni si sono impegnati

Log mining: towards a log-based Predictive Maintenance solutions for data centres

- Prototype of a log ingestion/management/processing system (for the INFN-CNAF use-case) on a Big Data Analytics platform towards a PM strategy, plus development of a clustering algorithm on logs based on similarity measures among textual fields to overcome limitations coming for the massive quantity of logs and their heterogeneity
- unsupervised Scholkopf's OCSVM, based on "volatility" observable from econophysics, applied to logs of a data center service (target: StoRM at CNAF Tier-1) → identify critical Δt in daily log and explore with NLP (namely, text processing) to extract high-level info
- Log parsing with deeper semantics interpretation of the anomalous occurrences, plus fuzzy-granular strategies (FBeM, eGNN, eGFC) in log analysis, to take advantage of the uncertain and imprecise nature of logs, yielding to propose a novel "evolving Log Parsing" (eLP) method for automatic parsing applied to system logs. In connection with CERN Operational Intelligence team.

VAE for Anomaly Detection

- Development of a joint variational autoencoder (JointVAE) ML model for AD in High-Energy Physics: once trained on known physics processes, it can identify anomalous events that correspond to previously unidentified physics signatures (w/o specific signature provided, i.e. fully unsupervised) + study of its implementation in latency-critical environments (e.g. CERN LHC triggers) and feasibility on FPGAs using a High-Level Synthesis (HLS) tool used in HEP named HLS4ML, which - combined with the quantisation of NNs - reduces model size, latency, energy consumption

Overview (in termini di lavori di **tesi** tematicamente rilevanti)

Lavori di tesi completati (non citati quelli in corso) selezionati in base a tematiche rilevanti per l'IG

- Legenda: LM + laurea magistrale, L = Laurea triennale, DSC PhD = dottorato in Data Science and Computation
- NB: non sono citati contributi di altri studenti (e.g. L. Giommi on MLaaS solutions) che hanno contribuito, ma le cui tesi non vertevano su tali argomenti come tematica specifica

L. Valente (LM Physics), "Joint Variational Auto-Encoder for Anomaly Detection in High Energy Physics, with an FPGA implementation"

L. Decker de Sousa (DSC PhD), "Data-stream driven Fuzzy-granular approaches for system maintenance"

S. Gasperini (LM Physics), "Exploiting Big Data solutions for CMS computing operations analytics"

F. Minarini (LM Physics), "Anomaly detection prototype for log-based predictive maintenance at INFN-CNAF Tier-1"

S. Rossi Tisbeni, "Big data analytics towards predictive maintenance at the INFN-CNAF computing centre"

G. Marastoni (L Physics), "Towards predictive maintenance at LHC computing centers: exploration of monitoring data at CNAF"

F. Minarini (L Physics), "CMS processing efficiency: Big Data exploration with Kibana and Elasticsearch"

S. De Luca (L Physics), "Studies of CMS data access patterns with machine learning techniques"

L. Ambroz (LM Physics), "Performance studies of CMS workflows using Big Data technologies"

Overview (in termini di articoli o proceedings)

Lista di contributi (selezionati), con ref accorciate per concisione

- "Joint Variational Auto-Encoder for Anomaly Detection in High Energy Physics", PoS ISGC&HEPiX2023 (2023) 014
- "Triggering Dark Showers with Conditional Dual Auto-Encoders", <https://arxiv.org/abs/2306.12955>
- "Explainable Log Parsing and Online Interval Granular Classification from Streams of Words", 2022 IEEE Int. Conf. Fuzzy Sys. (FUZZ-IEEE), 2022
- "Unsupervised Learning and Online Anomaly Detection: An On-Condition Log-Based Maintenance System", Int. Jour. of Embed. Real-Time Comm. Sys. (IJERTCS), Vol. 13, Issue 1, 2022
- "Exploiting Big Data solutions for CMS computing operations analytics", PoS ISGC2022 (2022) 006
- "Preparing Distributed Computing Operations for the HL-LHC Era With Operational Intelligence", Front.Big Data 4 (2022) 753409
- "A Big Data Platform for heterogeneous data collection and analysis in large-scale data centres", PoS ISGC2021 (2021) 008
- "Operational Intelligence for Distributed Computing Systems for Exascale Science", EPJ Web Conf. 245 (2020) 03017
- "Monitoring and Analytics at INFN Tier-1: the next step", EPJ Web Conf. 245 (2020) 07008
- "Real-Time Anomaly Detection in Data Centers for Log-based Predictive Maintenance using an Evolving Fuzzy-Rule-Based Approach", 2020 IEEE Int. Conf. Fuzzy Sys (FUZZ-IEEE), doi: 10.1109/FUZZ48607.2020.9177762
- "Comparison of Evolving Granular Classifiers applied to Anomaly Detection for Predictive Maintenance in Computing Centers", 2020 IEEE Conf. Evol. Adapt. Int. Sys. (EAIS), doi: 10.1109/EAIS48028.2020.9122779
- "Time-series anomaly detection applied to log-based diagnostic system using unsupervised machine learning approach", Conf of Open Innov. Assoc. (FRUCT), No. 27, 2020
- "Collection and harmonisation of system logs and prototypal Analytics services with the Elastic (ELK) suite at the INFN-CNAF computing centre", PoS ISGC2019 (2019) 027
- "Big Data Analysis for Predictive Maintenance at the INFN-CNAF Data Center using Machine Learning Approaches", Conf of Open Innov. Assoc. (FRUCT), 2019
- "Towards Predictive Maintenance with Machine Learning at the INFN-CNAF computing centre", PoS ISGC2019 (2019) 003
- "Progress on Machine and Deep Learning applications in CMS Computing", PoS ISGC2018 & FCDD (2018) 022
- "Progress in Machine Learning Studies for the CMS Computing Infrastructure", PoS ISGC2017 (2017) 023
- "Monitoring data transfer latency in CMS computing operations", J.Phys.Conf.Ser. 664 (2015) 3, 032033