EUROPEAN COMMISSION
DIRECTORATE-GENERAL
JOINT RESEARCH CENTRE
Directorate A – Strategy and Work Programme Coordination

CDP Theme

Machine Learning

Background on the JRC research

The interest of JRC in Machine Learning techniques (ML) has its roots in their applicative potential in dataset analysis, characterisation and clustering. Over the last decade, the JRC has exploited Machine Learning techniques mainly to:

1. provide data-driven decision making evidences to support the development of EU policies. For example to produce trend and forecast analysis which should be used to identify the best policy options for new regulations or recommendations support
2. Support applied research activities, e.g. ML as a part of broader research activities, enabling the comprehension of new phenomena using characterisation and clustering techniques.
3. Develop analysis tools in different domains, e.g. in the field of image recognition, text mining and analysis, network cyber-security, data protection, supply chain security, etc

JRC research on ML focuses on the following key research areas:

- *Cyber-Security and Data Protection*: JRC integrates ML in research activities in network data stream analysis, software analysis, verification and characterisation, privacy enhancing technologies design [1][2][3][4]. The techniques used in this area are mainly K-NN, Bayes Networks, Sequential Minimal Optimization, Support Vector Machine and Multilayer Perceptron neural network. The challenges, in these areas are mainly on the modelling and extraction of features and in the design of appropriate combination of ML approaches with strict time constraints.
- *Biometrics*: exploitation of ML techniques in support to law enforcement, mainly in the field of image recognition and sensor pattern noise [7][8]. The main families of ML techniques explored in this domain are Linear binary patterns, linear discriminant analysis, Adaboost and quadratic classifiers.
- *Context Based Retrieval and text analysis*: here, on a side, JRC interest and activities are oriented to the design of semantic space generators aiming at producing targeted wordlist for context based retrieval. On the other side, JRC designs and

exploits ML techniques for clustering and classification tasks in text analysis. An example of application is the exploitation of innovative SVM with polynomial kernels, K-nearest neighbours, Random Forests techniques for emotion classification and Recursive Neural Networks for text mining in large datasets [9][10]

- *Fight against counterfeiting*: identification and exploitation of the most suitable ML techniques to identify non-spoofable features of electronic devices [11][12].
- *Supply Chain Security and Customs Anti-Fraud:* JRC employs ML techniques in the analysis of billion-records size datasets of shipping containers traffic data in order to extract knowledge useful to the authorities in the supply chain security domain but also in the customs anti-fraud domain. Currently sequential pattern mining techniques (Conditional Random Fields) [13], Bayesian networks, decision trees, one-class classifiers [14] and spectrum clustering have been used to process the data. The on-going ML research focuses feature selection from sequential data, network analysis and change detection of probability distributions. The ongoing ML Research focuses feature selection from sequential data, network analysis and change detection of probability distributions.

Supported Policies

Machine learning, as research domain, cannot be linked directly to a single policy. It must be understood as an instrument used in diversified field of application to extract knowledge. As such machine learning supports indirectly potentially all the policy packages needing supporting evidences.

For example, considering the domains listed in the previous section, ML supports the following key policy initiatives:

- Implementation of the Digital Single Market strategy (incl. revision of the Privacy Directive)[1]
- Data Protection Package[2] [3]
- Big Data and free flow of data initiatives [4]
- Progress towards an effective and genuine Security Union[5]

Ongoing key projects and research

[1] "A Digital Single Market Strategy for Europe" COM(2015) 192 final

[2] DIRECTIVE (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA

[3] REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[4] "Towards a thriving data-driven economy". COM(2014) 442

[5] "Delivering on the European Agenda on Security to fight against terrorism and pave the way towards an effective and genuine Security Union" COM(2016)230 final

Machine Learning is a horizontal research domain which finds application in multiple areas also within the JRC. Here below some details on ongoing key JRC projects exploiting ML potential:

- ***Cyber-Security and Digital Identity*** **project (CSDI):** the project supports the EU policies in the area of security and digital single market with studies aiming at identifying threats undermining the cyber-security and privacy in the citizen digital life. Machine learning techniques are used here to identify new cyber-threats (software and malware analysis and classification), design mitigation techniques against cyber-attacks (Intrusion detection algorithms, antimalware analysis, assess online data leakages and design new, adaptable online identification protocols.
- ***Digital Forensic Investigation Techniques for Law Enforcement project (DFILE)*****:** the project aims at strengthening main critical investigation and prosecution functionalities such as identification, localization and crime content determination. ML, within the project, is a key element to analyse digital proofs.
- ***Competence Centre on Text Mining and Analysis for Policy, Security and Research***: Accurate, targeted, and timely information is needed by EU institutions at almost every stage of the decision making process. Text mining and analysis tools are necessary to address not only the problem of volume, but also of timeliness in order to provide the right information in the proper format for the decision making process, in a variety of contexts. This CC acts as a one stop shop for tools, services coaching and guidance on text mining issues for the EU Institutions and strategic partners.
- ***Container Traffic Analytics (ConTraffic):*** the project focuses on the analysis of the global traffic of shipping containers and the data processing techniques that can be used by authorities to better control and monitor these flows. The key goal is the intelligent exploitation of available data on the status and movement of shipping containers [25]. ML techniques are a key element of this data mining focused project.

Perspective direction of future JRC research interests in ML

While the domains of application described so far will evolve following the specific policy needs of the Commission' needs, it is however clear that from a scientific perspective JRC will need to enhance its competencies and capabilities in the area machine learning applied to big-data analysis with a specific emphasis on deep learning techniques. Areas where JRC may want to deepen its scientific base in the future include ML techniques for i) software analysis and software behavioural characterisation, ii) privacy and data protection evaluation, iii) image and biometric feature analysis and classification, iv) the detection of implicit forensic proofs, v) trajectory analysis of moving objects, vi) network analysis or deep learning techniques applied to network analysis and big data analysis; Furthermore, research on feature selection techniques from sequential data and on change detection of probability distributions are of interest for possible collaborations.

Over the past 5 years JRC scientists has successfully supervised several Ph.D. students on the topic of machine learning within the JRC in addition to external supervision of Ph.D. students in different universities.

Selected output for science and policy

Policy reports:

[16] “Fingerprint identification technology for its implementation in the Schengen Information System II (SIS-II)”, L. Beslay, J. Galbally, J.P. Nordvik, JRC Science for Policy Reports, 2015, EUR 27473, ISSN 1831-9424, ISBN 978-92-79-51929-1, DOI 10.2788/5062.

Peer reviewed papers:

[1] State-of-the-art in privacy preserving data mining. Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. 2004. SIGMOD Rec. 33, 1 (March 2004), 50-57. DOI=http://dx.doi.org/10.1145/974121.974131

[2] A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Bertino, E., Nai Fovino, I. & Provenza, L.P. Data Min Knowl Disc (2005) 11: 121. doi:10.1007/s10618-005-0006-6

[3] On the Efficacy of Static Features to Detect Malicious Applications in Android D Geneiatakis, R Satta, IN Fovino, R Neisse International Conference on Trust and Privacy in Digital Business, 87-98

[5] On the efficiency of user identification: a system-based approach. Malatras, A., Geneiatakis, D. & Vakalis, I. Int. J. Inf. Secur. (2016). doi:10.1007/s10207-016-0340-2

[7] 3D and 2.5D Face Recognition Spoofing Using 3D Printed Models. Galbally, J. & Satta, R. IET Biometrics, 2016, 5, 83-91

[8] Sensor Pattern Noise Matching Based on Reliability Map for Source Camera Identification. VISAPP (1) 2015: 222-226

[9] Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. A Balahur, M Turchi. Computer Speech & Language, 2014.

[10] Sentiment analysis system adaptation for multilingual processing: The case of tweets . A Balahur, JM Perea-Ortega. Information Processing & Management, 2015

[11] “Experimental identification of smartphones using fingerprints of built-in microelectro mechanical systems (mems),” G. Baldini, G. Steri, F. Dimc, R. Giuliani, and R. Kamnik, . Sensors, vol. 16, no. 6, p. 818, 2016.

[12] GNSS Receiver Fingerprinting for Security-Enhanced Applications. Daniele Borio, Ciro Gioia, Gianmarco Baldini and Joaquin Fortuny, European Commission, Joint Research Centre (JRC), Directorate E – Space, Security & Migration, Italy. Proceedings of ION GNSS+, Portland, Oregon, USA, 2016

[13] Inferring itineraries of containerized cargo through the application of conditional random fields. P. Chahuara, L. Mazzola, M. Makridis, C. Schifanella, A. Tsois, and M. Pedone. In Proc. of the 2014 IEEE Joint Intelligence and Security Informatics Conf., pages 137--144, 2014, doi:10.1109/JISIC.2014.29

[14] Detecting anomalous maritime container itineraries for anti-fraud and supply chain security. E. Camossi, T. Dimitrova, and A. Tsois. In Proc. of the 2012 European Intelligence and Security Informatics Conf., EISIC '12, pages 76-83, 2012.

[15] A. Tsois, J. A. Cotelo Lema, M. Makridis, and E. Checchi. Using container status messages to improve targeting of high-risk cargo containers. In Research Track at the 5th World Customs Organization Technology & Innovation Forum, Rotterdam, Netherlands, 2015

[17] “Mobile App anonymization control”, I. Nai Fovino, G. Stery, R. Neisse, R. Satta. JRC Technical Report, 2015

Hosting Directorate

Directorate: Space, Security and Migration (DIR E)