algorithms for the clustering of features and the extraction of association rules.

The basic ML algorithms used in this platform are:
- Expectation-Maximisation algorithm (EM): Decides the optimal number of clusters.
- K-means algorithm: Identifies different groups of users and opinions.
- Apriori algorithm: Determines interesting and useful relationships among attributes.

Our methodology implements five different information needs emerging from users throughout the sense-making process.
- Same profiles with similar opinions: Identifies different user profiles that share similar opinions on specific positions.
- Sharing similar opinions with specific user: Extracts different groups of users whose opinions are closely related to specific user(s).
- Same users with similar preferences: Determines groups of users who share both similar opinions and profile characteristics.
- Similar opinions based on different profiles: Identifies similar opinions expressed by users with different profile characteristics.
- Different user profiles with similar/dissimilar opinions: Finds different groups of user profiles who share similar or dissimilar opinions.

There are several avenues that are worthy of further investigation, with an emphasis on improving the system's usability. Thorough evaluations with real users and large datasets of discussions will be carried out, including expert walk-through evaluation of the platform and adjustments based on user expert feedback.

Links:

[L1] www.ics.forth.gr/isl/apopsis
[L2] http://www.ics.forth.gr/isl/mace/mace.rdfs
[L3] http://www.ics.forth.gr/isl/mace/MACE%20Ontology_Scope_Notes.pdf

References:

[1] J. Conklin, M.L. Begeman: "gIBIS: A hypertext tool for team design deliberation", Proc. of the ACM conference on Hypertext, ACM, 1987.
[2] J.Schneider, T.Groza, and A.Passant: "A review of argumentation for the social semantic web", Semantic Web 4.2 (2013)
[3] T.Patkos, G.Flouris, and A.Bikakis: "Symmetric Multi-Aspect Evaluation of Comments", ECAI 2016-22nd European Conference on Artificial Intelligence, 2016.

Please contact:

Elisjana Ymeralli, FORTH, Greece
ymeralli@ics.forth.gr (mailto:ymeralli@ics.forth.gr)

Can we Trust Machine Learning Results? Artificial Intelligence in Safety-Critical Decision Support

by Katharina Holzinger (SBA Research), Klaus Mak, (Austrian Army) Peter Kieseberg (St. Pölten University of Applied Sciences), Andreas Holzinger (Medical University Graz, Austria)

Machine learning has yielded impressive results over the last decade, but one important question that remains to be answered is: How can we explain these processes and algorithms in order to make the results applicable as proof in court?

Artificial intelligence (AI) and machine learning (ML) are making impressive impacts in a range of areas, such as speech recognition, recommender systems, and self-driving cars. Amazingly, recent deep learning algorithms, trained on extremely large data sets, have even exceeded human performance in visual tasks, particularly on playing games such as Atari Space Invaders, or mastering the game of Go [L1]. An impressive example from the medical domain is the recent work by Esteva et al. (2017) [1]: they utilised a GoogleNet Inception v3 convolutional neural network (CNN) architecture for the classification of skin lesions, pre-trained their network with approximately 1.3 million images (1,000 object categories), and trained it on nearly 130,000 clinical images. The performance was tested against 21 board-certified dermatologists on biopsy-proven clinical images. The results show that deep learning can achieve a performance on par with human experts.

One problem with such deep learning models is that they are considered to be "black-boxes", lacking explicit declarative knowledge representation, hence they have difficulty in generating the required underlying explanatory structures. This is limiting the achievement of their full potential, and even if we understand the mathematical theories behind the machine model, it is still complicated to get insight into the internal workings. Black box models lack transparency and one question is becoming increasingly important: "Can we trust the results?" We argue that this question needs to be rephrased into: "Can we explain how and why a result was achieved?" (see Figure 1). A classic example is the question "Which objects are similar?" This question is the typical pretext for using classifiers to classify data objects, e.g. pictures, into different categories (e.g., people or weapons), based on utilising implicit models derived through training data. Still, an even more interesting question, both from theoretical and legal points of view, is "Why are those objects similar?" This is especially important in situations when the results of AI are not easy to verify. While verification of single items is easy enough in many classical problems solved with machine learning, e.g., detection of weapons in pictures, it can be a problem in cases where the verification cannot be

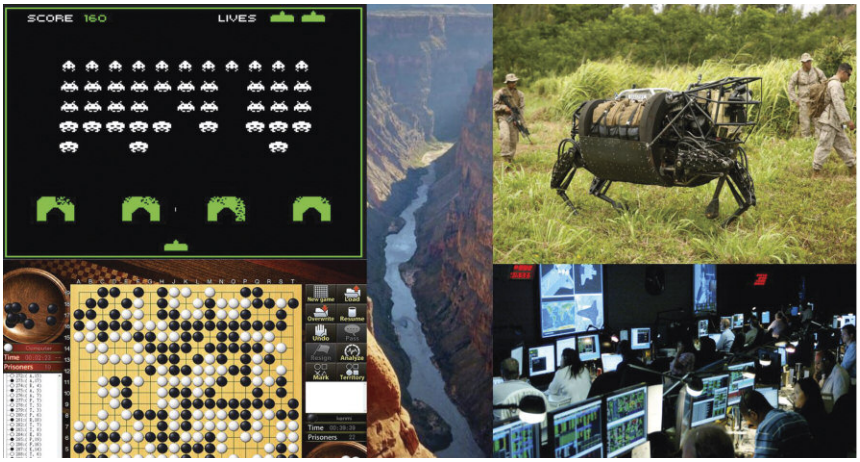

Figure 1: AI shows impressive success, still having difficulty in generating underlying explanatory structures.

done by a human with (close to) 100 percent precision, as in cancer detection, for example.

Consequently, there is growing demand for AI, which not only performs well, but is also transparent, interpretable and trustworthy. In our recent research, we have been working on methods and models to reenact the machine decision-making process [2], to reproduce and to comprehend the learning and knowledge extraction processes, because for decision support it is very important to understand the causality of learned representations. If human intelligence is complemented by machine learning, and in some cases even overruled, humans must be able to understand, and most of all to be able to interactively influence the machine decision process. This needs sense making to close the gap between human thinking and machine "thinking".

This is especially important when the algorithms are used to extract evidence from data to be used in court. A similar discussion has already been started in the area of digital forensics, with novel forensic processes being able to extract small parts of information from all over the IT-system in question and then combining them in order to generate evidence – which is currently not usable in court, simply owing to the fact that no judge will rule on evidence gathered by a process no one can control and see through.

Our approach will also address rising legal and privacy concerns, e.g., with the new European General Data Protection Regulation (GDPR and ISO/IEC 27001) entering into force on May, 25, 2018. This regulation will make black-box approaches difficult to use in business, because they are not able to explain why a decision has been made. In addition, it must be noted that the "right to be forgotten" [3] established by the European Court of Justice has been extended to become a "right of erasure"; it will no longer be sufficient to remove a person's data from search results when requested to do so, data controllers must now erase that data. This is especially problematic when data is used inside the knowledge base of a machine learning algorithm, as changes here might make decisions taken and results calculated by the algorithm irreproducible. Thus, in order to understand the impact of changes to such results, it is of vital importance to understand the internals of the intrinsic model built by these algorithms. Furthermore, and in spirit with our second research interest in this area, the deletion of data is a highly complicated process in most modern complex environments and gets even more complicated when considering the typical targets of data provisioning environments like databases that are opposing deletion:

- Fast searches: Typically, one main goal in database design is to provide fast and efficient data retrieval. Thus, the internal structures of such systems have been designed in order to speed up the search process by incorporating parts of the content as search keys, yielding a tree structure that is organised along the distribution of key information, thus making deletion a problematic issue.
- Fast data manipulation: Like in modern file systems, data entries that are "deleted" are not actually erased from the disk with overwriting the respective memory for performance reasons, but only unlinked from the search indices and marked for overwriting.
- Crash Recovery: Databases must possess mechanisms in case an operation fails (e.g., due to lack of disk space) or the database crashes in the middle of a data-altering operation (e.g., blackouts) by reverting back to a consistent state that is throughout all data tables. In order to provide this feature, transaction mechanisms must store the data already written to the database in a crash-safe manner, which can be used, for example, in forensic investigations to uncover deleted information.
- Data Replication: Companies have implemented mirror data centres that contain replicated versions of the operative database in order to be safe against failures. Deletion from such systems is thus especially complicated.

With our research, we will be able to generate mechanisms for better understanding and control of the internal models of machine learning algorithms, allowing us to apply fine-grained changes on one hand, and to better estimate the impact of changes in knowledge bases on the other hand. In addition, our research will yield methods for enforcing the right to be forgotten in complex data driven environments.

Link: [L1] https://arxiv.org/abs/1708.01104

References:

[1] A. Esteva, et al.: "Dermatologist-level classification of skin cancer with deep neural networks", Nature, 542, (7639), 115-118, 2017, doi:10.1038/nature21056.

[2] A. Holzinger, et al.: "A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop", arXiv:1708.01104.

[3] B. Malle, P. Kieseberg, S. Schrittwieser, A. Holzinger: "Privacy Aware Machine Learning and the 'Right to be Forgotten'", ERCIM News 107, (3), 22-23, 2016.

Please contact:

Katharina Holzinger, SBA Research, Vienna, Austria
kholzinger@sba-research.org