Towards interactive Machine Learning for solving complex problems in Health Informatics

Andreas Holzinger
Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation
Medical University Graz, Austria
Institute of Information Systems and Computer Media, Graz University of Technology, Austria
a.holzinger@hci.kdd.org

September, 27, 2016
KI-Workshop, Klagenfurt, Austria

Abstract

In automatic machine learning (aML) great advances have been made, for example, in speech recognition, recommender systems, or autonomous vehicles. Automatic approaches greatly benefit from big data with many training sets. However, in health informatics, we are often confronted with a small number of data sets or rare events, where aML suffer of insufficient training samples. Here interactive Machine Learning (iML) may be of help, having its roots in reinforcement learning, preference learning and active learning. The term iML is not yet well used, so we define it as algorithms that can interact with agents and can optimize their learning behavior through these interactions, where the agents can also be human. This human-in-the-loop" can be beneficial in solving computationally hard problems, e.g., subspace clustering, protein folding, or k-anonymization of health data, where human expertise can help to reduce an exponential search space through heuristic selection of samples. Therefore, what would otherwise be an NP-hard problem, reduces greatly in complexity through the input and the assistance of a human agent involved in the learning phase. For the successful application of ML in health informatics a multidisciplinary skill set is required, encompassing the following seven specializations: 1) data science, 2) algorithms, 3) network science, 4) graphs/topology, 5) time/entropy, 6) data visualization, and 7) privacy, data protection, safety and security, fostered in the HCI-KDD approach. After giving a very brief introduction to the HCI-KDD approach, I start this talk with showing some problems with probabilistic information in the health domain. After the definition of aML and showing some examples I will provide some insight into our latest iML-research.

Extended Abstract

The original idea of the HCI-KDD approach [1] is in combining aspects of the best of two worlds: Human–Computer Interaction (HCI), with emphasis on cognitive science, particularly dealing with *human intelligence*, and Knowledge Discovery/Data Mining (KDD), with emphasis on machine learning, particularly dealing with *computational intelligence* [2]. Cognitive science (CS) studies the principles of human learning to understand intelligence. In CS our natural surrounding is in $\mathbb{R}^3$ and it is amazing how humans extract so much from so little in dimensions of ≤ 3. However, the problem in health informatics is that we are challenged with data of arbitrarily high dimensions [3], [4]. Within such data, relevant *structural* patterns and/or *temporal* patterns ("knowledge") are often hidden, difficult to extract, hence not accessible to a biomedical expert. The challenge is to bring the

results from high dimensions into the lower dimension, because the health experts are working on 2D surfaces. CS and Machine Learning (ML) did not harmonize in the past: Whilst CS had its focus on specific experimental paradigms because it was embedded deeply in Psychology and Linguistics and aimed to be cognitively/neutrally plausible, ML had its focus on standard learning problems and tried to optimize in the range of 1 % because it was embedded in Computer Engineering, and aimed to have working systems to solve practical problems whether mimicking the human brain or not. The ultimate goal ever since is to develop algorithms which can *automatically* learn from data, hence can improve with experience over time. Sometimes the application of such aML approaches in the complex health domain fails, and a good example are Gaussian processes, where aML approaches (e.g. standard kernel machines) struggle on function extrapolation problems which are trivial for human learners [5]. Consequently, iML-approaches can be of interest to solve problems, where we are lacking big data sets, deal with complex data and/or rare events, where traditional learning algorithms suffer due to insufficient training samples [6]. Here the doctor-in-the-loop [7] can help, where human expertise and long-term experience can assist in solving problems which otherwise would remain NP-hard; examples include subspace clustering [8], protein folding, or privacy preserving ML, which is an important issue, fostered by anonymization, in which a record is released only if it is indistinguishable from k other entities in the data, but where k-anonymity is highly dependent on spatial locality in order to effectively implement the technique in a statistically robust way. In high dimensionalities data becomes sparse, hence the concept of spatial locality is not easy to define. Consequently, it becomes difficult to anonymize the data without an unacceptably high amount of information loss - here iML could be of help [9].

Much future research has to be done, particularly in the fields of Multi-Task Learning and Transfer Learning and to go towards Multi-Agent-Hybrid Systems. As example I will finally discuss a recent project on tumor growth: The underlying complexity of cancer demands for abstractions to disclose an exclusive subset of information related to this disease. Machine learning for tumor growth profiles and model validation may be of great help here, where we follow the assumption that the key to understanding the concepts of cancer lies within an integrative translation and multi-dimensional connection of open data sets [10], [11].

Graphical Abstract

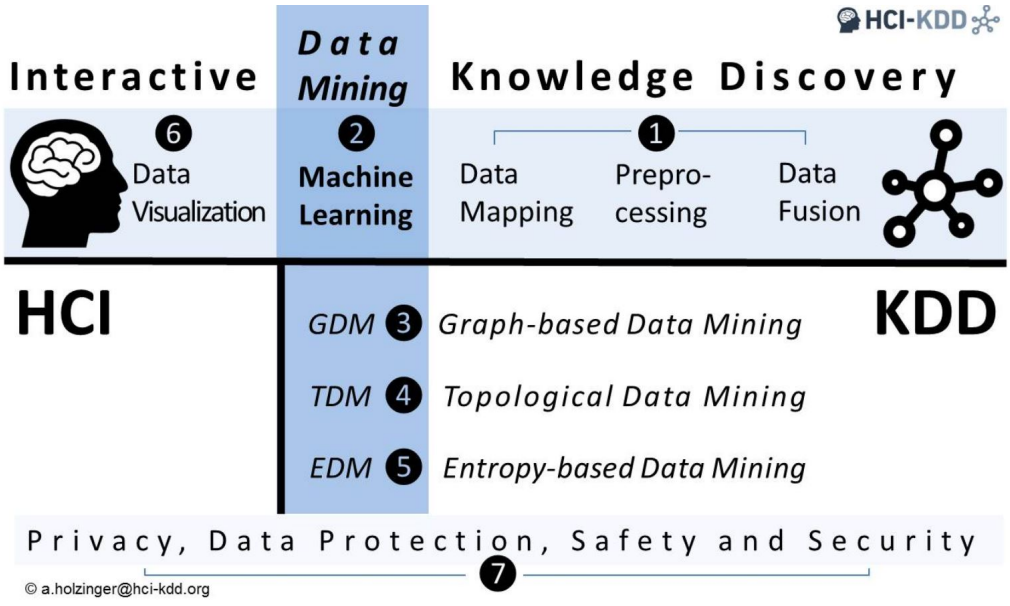

Figure 1: Graphical abstract: The interactive knowledge discovery pipeline, ML is at the heart, but solving problems in health informatics needs a concerted effort of several other disciplines (Image taken from hci-kdd.org)

Short Bio
Andreas Holzinger, lead of the Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation at the Medical University Graz, Associate Professor (Univ.-Doz.) for Applied Computer Science at the Institute of Information Systems and Computer Media at Graz University of Technology. Currently, Andreas is Visiting Professor for Machine Learning in Health Informatics [12], at the Faculty of Informatics at Vienna University of Technology. His research interests are in supporting human intelligence with machine learning to help to solve problems in the biomedical domain. Andreas obtained a Ph.D. in Cognitive Science from Graz University in 1998 and his Habilitation (second Ph.D.) in Applied Computer Science from Graz University of Technology in 2003. Andreas was Visiting Professor in Berlin, Innsbruck, London (2 times), and Aachen. Andreas founded the international Expert Network HCI-KDD to foster a synergistic combination of methodologies of two areas that offer ideal conditions towards unraveling problems in understanding complex data: HumanComputer Interaction (HCI) and Knowledge Discovery from Data (KDD), with the goal of supporting human intelligence with machine intelligence for knowledge discovery. Andreas is Associate Editor of Knowledge and Information Systems (KAIS), and member of IFIP WG 12.9 Computational Intelligence. h-Index=37, 6933 citations (15.8.2016), http://hci-kdd.org/

References

[1] Andreas Holzinger. Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. *IEEE Intelligent Informatics Bulletin*, 15(1):6–14, 2014.

[2] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Lecture Notes in Computer Science LNCS 8401*, pages 1–18. Springer, Heidelberg, Berlin, 2014.

[3] Andreas Holzinger, Matthias Dehmer, and Igor Jurisica. Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. *BMC Bioinformatics*, 15(S6):I1, 2014.

[4] Sangkyun Lee and Andreas Holzinger. Knowledge discovery from complex high dimensional data. In *Solving Large Scale Learning Tasks. Challenges and Algorithms, Lecture Notes in Artificial Intelligence, LNAI 9580*. 2016.

[5] Andreas Holzinger. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics)*, 3(2):119–131, 2016.

[6] Andreas Holzinger. Interactive machine learning (iML). *Informatik Spektrum*, 39(1):64–68, 2016.

[7] Dominic Girardi, Josef Kueng, Raimund Kleiser, Michael Sonnberger, Doris Csillag, Johannes Trenkler, and Andreas Holzinger. Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research. *Brain Informatics*, pages 1–11, 2016.

[8] Michael Hund, Dominic Bhm, Werner Sturm, Michael Sedlmair, Tobias Schreck, Torsten Ullrich, Daniel A. Keim, Ljiljana Majnaric, and Andreas Holzinger. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the doctor-in-the-loop. *Brain Informatics*, pages 1–15, 2016.

[9] Bernd Malle, Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. The right to be forgotten: Towards machine learning on perturbed knowledge bases. In *Springer Lecture Notes in Computer Science LNCS 9817*, pages 251–256. Springer, Heidelberg, Berlin, New York, 2016.

[10] Fleur Jeanquartier, Claire Jean-Quartier, Tobias Schreck, David Cemernek, and Andreas Holzinger. Integrating open data on cancer in support to tumor growth analysis. In *Lecture Notes in Computer Science LNCS 9832*. 2016.

[11] Fleur Jeanquartier, Claire Jean-Quartier, David Cemernek, and Andreas Holzinger. In silico modeling for tumor growth visualization. *BMC Systems Biology*, 10(1):1–15, 2016.

[12] Andreas Holzinger. Machine learning for health informatics. 2016. URL http://hci-kdd.org/machine-learning-for-health-informatics-course/.