5-Page Personal Research Statement: Machine Learning & Knowledge Extraction (MAKE) for solving problems in health informatics

Andreas Holzinger
Holzinger Group HCI-KDD, Institute for Medical Informatics/Statistics
Medical University Graz, Austria
andreas.holzinger@medunigraz.at
Graz, April, 2018

Abstract

Together with my group and in a joint effort with my international colleagues, I am interested in theoretical, algorithmic, and experimental studies in machine learning in order to solve the problem of knowledge extraction from complex data to discover unknown unknowns for solving problems in health informatics. I am excited to help to answer a grand question: *How can we perform a new task by exploiting knowledge extracted during problem solving of previous tasks.* Contributions to this problem would have major impact on AI generally, and Machine Learning (ML) specifically, as we could solve a lot of problems in health informatics. Ultimately, to reach a level of *usable* intelligence, we need 1) to learn from prior data, 2) to extract knowledge, 3) to generalize - i.e. guessing where probability mass/density concentrates, 4) to fight the curse of dimensionality, and 5) to disentangle underlying *explanatory factors*, i.e. to *make sense* of the data in the *context* of the medical problem. However, the application of automatic machine learning (aML) in health informatics has some drawbacks. A good example are Gaussian processes, where aML (e.g. kernel machines) struggle on function extrapolation problems, which are trivial for human experts. Consequently, interactive machine learning (iML) with a human-in-the-loop, thereby making use of human cognitive abilities, can be of particular interest to solve problems, where learning algorithms suffer of lacking training samples, dealing with complex data and/or rare events or computationally hard problems, e.g. subspace clustering, protein folding, or k-anonymization. One increasingly important aspect is that automatic methods are considered as black boxes, and raising legal and privacy issues in the European Union make "glass box" approaches important in the future to be able to make decisions transparent, re-traceable, understandable, thus explainable. This helps to explain *why* a machine decision has been made, paving the way towards explainable AI.

1 Introduction and Motivation

Machine learning (ML) as a field started seven decades ago with ideas on developing algorithms that can automatically learn from data to gain knowledge from experience and to gradually improve their learning behaviour. The field revolved at the intersection of cognitive science and computer science [1], and progressed enormously in the last two decades with huge application challenges and business potential. The best practice examples today are autonomous vehicles, recommender systems, or natural language understanding [2], [3]. To see health informatics among the greatest challenges is not surprising, because here we are confronted with uncertainty, with probabilistic, unknown, incomplete, heterogenous, noisy, dirty, erroneous, inaccurate and missing data sets in arbitrarily high dimensional spaces [4], [5].

Consequently, to apply ML successfully for solving problems in health informatics, a concerted cross-domain effort is required bringing together experts from four main areas in a catalytic way: 1) data science, 2) learning algorithms, 3) visualization/visual analytics, and 4) privacy, data protection, safety and security. Only in a joint effort we can reach a level of **useable intelligence,** and bringing AI/ML directly into the work-flows of the end users (refer to the HCI-KDD approach towards integrative machine learning [6]).

My contributions to the international research community are three-fold, i.e.:
1) Contributing to the international research community (example: `https://goo.gl/gXnu04` on the design, development and testing of novel methods, i.e. interactive machine learning (iML) with the human-in-the-loop to help to solve problems in health informatics, fostering open source, open data and open access.
2) Building and maintaining an international network of experts with complementary interests but sharing a common goal, and organizing workshops and a flagship conference: `https://cd-make.net`
3) Teaching and mentoring Machine Learning & Knowledge Extraction with application in health informatics at various levels: `https://youtu.be/lc2hvuh0FwQ` and see teaching statement: `https://goo.gl/Abv0gc`.

2 Automatic Machine Learning vs. interactive Machine Learning

Let us consider n data contained in a set $\mathcal{D} = x_{1:n} = \{x_1, x_2, ..., x_n\}$, let be the likelihood $p(\mathcal{D}|\theta)$ and specify a prior $p(\theta)$, consequently we can compute the posterior:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

The inverse probability allows us to learn from data, infer unknowns and to make predictions. However, the performance of any ML algorithm is dependent on the choice of the data representations. These features are key for learning and understanding, hence much effort in ML goes into the design of preprocessing pipelines and in data transformations and data mappings that result in a respective representation which supports effective machine learning. Current learning algorithms have still an enormous weakness: they are unable to *extract the discriminative knowledge* from the data. Consequently, it is of utmost importance for us, to expand the applicability of learning algorithms, hence, to make them less dependent on feature engineering (a new kind of algorithm usability). A truly intelligent algorithm must be able to understand the *context.* [7] argue that this can only be achieved if the algorithms can learn to identify and disentangle the underlying exploratory factors already existent among the low-level data.

The ultimate goal is to develop algorithms which can *automatically* learn from data, hence can improve with experience over time *without any human-in-the-loop.* [8]. Such approaches work well when having large amounts of data [9]. However, the application of such aML-approaches in complex domains - such as health - seems elusive in the near future 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. Consequently, iML-approaches, by integrating a human-into-the-loop (e.g. a human kernel [10], or the involvement of a human directly into a machine-learning algorithm [1] thereby making use of human cognitive abilities, is a promising approach. iML-approaches can be of particular 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. Here the "doctor-in-the-loop" can help, where human expertise and human experience can assist in solving problems which otherwise would remain NP-hard.

We focused in a recent work [11] on the Traveling Salesman Problem (TSP). This appears in a number of practical problems in health informatics, e.g. the native folded three-dimensional conformation of a protein is its lowest free energy state and both 2D and 3D folding processes as a free energy minimization problem belong to a large set of computational problems, assumed to be very hard ("conditionally intractable") [12]. The TSP basically is about finding the shortest path through a set of points, returning to the origin. As it is an intransigent mathematical problem, many heuristics have been developed in the past to find approximate solutions [13]. Most of all this helps to go towards explainable AI [14].

[1] see a recent experiment: https://hci-kdd.org/gamification-interactive-machine-learning

3 Future Challenges

Multi-task learning (MTL) aims to improve the prediction performance by learning a problem together with multiple, different but related other problems through shared parameters or a shared representation. The underlying principle is *bias learning* based on probable approximately correct learning (PAC learning) [15]. To find such a bias is still the hardest problem in any ML task and essential for the initial choice of an appropriate hypothesis space, which must be large enough to contain a solution, and small enough to ensure a good generalization from a small number of data sets. Existing methods of bias generally require the input of a human-expert-in-the-loop in the form of heuristics and domain knowledge to ensure the selection of an appropriate set of features, as such features are key to learning and understanding. However, such methods are limited by the accuracy and reliability of the expert' s knowledge (robustness of the human) and also by the extent to which that knowledge can be transferred to new tasks (see next subsection). Baxter (2000)[16] introduced a model of bias learning which builds on the PAC learning model which concludes that learning multiple related tasks reduces the sampling burden required for good generalization and bias that is learnt on sufficiently many training tasks is likely to be good for learning novel tasks drawn from the same environment (the problem of transfer learning to new environments is discussed in the next subsection). A practical example is *regularized MTL* [17], which is based on the minimization of regularization functionals similar to Support Vector Machines (SVMs), that have been successfully used in the past for single–task learning. The regularized MTL approach allows to model the relation between tasks in terms of a novel kernel function that uses a task–coupling parameter and largely outperforms single–task learning using SVMs. However, multi-task SVMs are inherently restricted by the fact that SVMs require each class to be addressed explicitly with its own weight vector. In a multi-task setting this requires the different learning tasks to share the *same set of classes.* An alternative formulation for MTL is an extension of the large margin nearest neighbor algorithm (LMNN) [18]. Instead of relying on separating hyper-planes, its decision function is based on the nearest neighbor rule which inherently extends to many classes and becomes a natural fit for MTL. This approach outperforms state-of-the-art MTL classifiers, and here many research challenges remain open which I want to attack [19].

Transfer learning is the ability to learn tasks permanently and this is crucial to the development of any artificial intelligence. Humans can do that very good - even very little children. A good counterexample are neural networks (deep learning) which in general are not capable of it and are considerably hampered by *catastrophic forgetting*.

The synaptic consolidation in human brains enables continual learning by reducing the plasticity of synapses that are vital to previously learned tasks. [20] implemented an algorithm that performs a similar operation in artificial neural networks by constraining important parameters to stay close to their old values. As known a deep neural network consists of multiple layers of linear projections followed by element-wise non-linearities. Learning a task consists basically of adjusting the set of weights and biases θ of the linear projections, consequently, many configurations of θ will result in the same performance which is relevant for the so-called elastic weight consolidation (EWC): over-parametrization makes it likely that there is a solution for task B, θ^*_B, that is close to the previously found solution for task A, θ^*_A. While learning task B, EWC therefore protects the performance in task A by constraining the parameters to stay in a region of low error for task A centered around θ^*_A. This constraint has been implemented as a quadratic penalty, and can therefore be imagined as a mechanical spring anchoring the parameters to the previous solution, hence the name elastic.

In order to justify this choice of constraint and to define which weights are most important for a task, it is useful to consider neural network training from a probabilistic perspective. From this point of view, optimizing the parameters is tantamount to finding their most probable values given some data $\mathcal{D}$. Interestingly, this can be computed as conditional probability $p(\theta|\mathcal{D})$ from the prior probability of the parameters $p(\theta)$ and the probability of the data $p(\mathcal{D}|\theta)$ by:

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D})$$

Here I want to contribute on avoiding the problem of catastrophic forgetting, which is a hot topic with many open research avenues [21].

According to Pan & Yang (2010) [22] a major assumption in many ML algorithms is, that both the training data and future (unknown) data must be in the same feature space and required to have the

same distribution. In many real-world applications, particularly in the health domain, this is not the case: Sometimes we have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter data may be in a completely different feature space or follows a different data distribution. In such cases transfer learning would greatly improve the performance of learning by avoiding much expensive data-labeling efforts, however, many open questions remain for future research [23].

Multi-Agent-Systems (MAS) are collections of many agents interacting with each other. They can either share a common goal (for example an ant colony, bird flock, or fish swarm etc.), or they can pursue their own interests (for example as in an open-market economy). MAS can be traditionally characterized by the facts that (a) each agent has incomplete information and/or capabilities for solving a problem, (b) agents are autonomous, so there is no global system control; (c) data is decentralized; and (d) computation is asynchronous [24]. For the health domain of particular interest is the *consensus problem*, which formed the foundation for distributed computing [25]. The roots are in the study of (human) experts in group consensus problems: Consider a group of humans who must act together as a team and each individual has a subjective probability distribution for the unknown value of some parameter; a model which describes how the group reaches agreement by pooling their individual opinions was described by DeGroot (1974)[26] and was used decades later for the aggregation of information with uncertainty obtained from multiple sensors [27] and medical experts [28]. On this basis Olfati-Saber et al. (2007)[29] presented a theoretical framework for analysis of consensus algorithms for networked multi-agent systems with fixed or dynamic topology and directed information flow. In complex real-world problems, e.g., for the epidemiological and ecological analysis of infectious diseases, standard models based on differential equations very rapidly become unmanageable due to too many parameters, and here MAS can also be very helpful [30]. Moreover, collaborative multi-agent reinforcement learning has a lot of research potential for machine learning [31]. Here we did some preliminary work on collaborative interactive machine learning where I want to contribute with future work.

Acknowledgments

I am very grateful for feedback and long-term support from my international research colleagues.

References

[1] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. doi: 10.1126/science.1192788.

[2] Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. URL `http://dx.doi.org/10.1126/science.aaa8415`.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. URL `http://dx.doi.org/10.1038/nature14539`.

[4] 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. doi: 10.1186/1471-2105-15-S6-I1.

[5] Sangkyun Lee and Andreas Holzinger. Knowledge discovery from complex high dimensional data. In Stefan Michaelis, Nico Piatkowski, and Marco Stolpe, editors, *Solving Large Scale Learning Tasks. Challenges and Algorithms, Lecture Notes in Artificial Intelligence, LNAI 9580*, pages 148–167. Springer, Cham, 2016. URL `http://dx.doi.org/10.1007/978-3-319-41706-6_7`.

[6] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In Andreas Holzinger and Igor Jurisica, editors, *Lecture Notes in Computer Science LNCS 8401*, pages 1–18. Springer, Heidelberg, 2014. URL `http://dx.doi.org/10.1007/978-3-662-43968-5_1`.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. URL `http://dx.doi.org/10.1109/TPAMI.2013.50`.

[8] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

[9] Soeren Sonnenburg, Gunnar Raetsch, Christin Schaefer, and Bernhard Schoelkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(7):1531–1565, 2006. URL `http://www.jmlr.`

org/papers/v7/sonnenburg06a.html.

[10] Andrew G. Wilson, Christoph Dann, Chris Lucas, and Eric P. Xing. The human kernel. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems, NIPS 2015*, volume 28, pages 2836–2844, 2015.

[11] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *Springer Lecture Notes in Computer Science LNCS 9817*, pages 81–95. Springer, Heidelberg, Berlin, New York, 2016. doi: 10.1007/978-3-319-45507-56.

[12] Pierluigi Crescenzi, Deborah Goldman, Christos Papadimitriou, Antonio Piccolboni, and Mihalis Yannakakis. On the complexity of protein folding. *Journal of computational biology*, 5(3):423–465, 1998. doi: 10.1016/S0092-8240(05)80170-3.

[13] J. N. Macgregor and T. Ormerod. Human performance on the traveling salesman problem. *Perception & Psychophysics*, 58(4):527–539, 1996. doi: 10.3758/bf03213088.

[14] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable ai systems for the medical domain? *arXiv:1712.09923*, 2017.

[15] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. URL http://dx.doi.org/10.1145/1968.1972.

[16] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research (JAIR)*, 12: 149–198, 2000. URL http://dx.doi.org/10.1613/jair.731.

[17] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004. URL http://dx.doi.org/10.1145/1014052.1014067.

[18] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009. URL http://www.jmlr.org/papers/v10/weinberger09a.html.

[19] Shibin Parameswaran and Kilian Q Weinberger. Large margin multi-task metric learning. In John Lafferty, Christopher Williams, John Shawe-Taylor, Richard Zemel, and Aron Culotta, editors, *Advances in neural information processing systems 23 (NIPS 2010)*, pages 1867–1875, 2010. URL http://papers.nips.cc/paper/3935-large-margin-multi-task-metric-learning.

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv preprint arXiv:1612.00796*, 2016. URL https://arxiv.org/abs/1612.00796.

[21] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. *arXiv preprint*, 2015. URL https://arxiv.org/abs/1312.6211v3.

[22] Sinno J. Pan and Qiang A. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. URL http://dx.doi.org/10.1109/tkde.2009.191.

[23] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685, 2009.

[24] Katia P Sycara. Multiagent systems. *AI magazine*, 19(2):79, 1998. URL http://aitopics.org/topic/multi-agent-systems.

[25] Nancy A Lynch. *Distributed algorithms*. Morgan Kaufmann, San Francisco, 1996.

[26] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974.

[27] Jon A Benediktsson and Philip H Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man and Cybernetics*, 22(4):688–704, 1992.

[28] Susan C Weller and N Clay Mann. Assessing rater performance without a gold standard using consensus theory. *Medical Decision Making*, 17(1):71–79, 1997.

[29] Reza Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007. doi: 10.1109/jproc.2006.887293.

[30] B. Roche, J. F. Guegan, and F. Bousquet. Multi-agent systems in epidemiology: a first step for computational biology in the study of vector-borne disease transmission. *BMC Bioinformatics*, 9, 2008. doi: 10.1186/1471-2105-9-435.

[31] Jelle R Kok and Nikos Vlassis. Collaborative multiagent reinforcement learning by payoff propagation. *The Journal of Machine Learning Research*, 7:1789–1828, 2006.