Towards Interactive Recommender Systems with the Doctor-in-the-Loop

Andreas Holzinger[1], André Calero Valdez[1,2], Martina Ziefle[2]

Holzinger Group, HCI-KDD, Institute for Medical Informatics, Medical University Graz[1]
Human-Computer Interaction Center, RWTH-Aachen University[2]

Abstract

Recommender Systems are a perfect example for automatic Machine Learning (aML) – which is the fastest growing field in computer science generally and health informatics specifically. The general goal of ML is to develop algorithms which can learn and improve over time and can be used for predictions and decision support – which is of the central interest of health informatics. Whilst automatic approaches greatly benefit from big data with many training sets, in the health domain experts are often confronted with a small number of complex data sets or rare events, where aML-approaches suffer of insufficient training samples. Here interactive Machine Learning (iML) may be of help, which can be defined as "algorithms that can interact with agents and can optimize their learning behaviour through these interactions, where the agents can also be human". Such a human can be an expert, i.e. a medical doctor, and this "doctor-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 expert agent involved in the learning phase. Important future research aspects are in the combined use of both human intelligence and computer intelligence, in the context of hybrid multi-agent recommender systems which can also make use of the power of crowdsourcing to make use of joint decision making – which can be very helpful e.g. in the diagnosis and treatment of rare diseases.

1 Introduction

Originally the term machine learning (ML) was defined as *"... artificial generation of knowledge from experience",* and the first studies have been performed with games, i.e., with the game of checkers (Samuel, 1959). Today, ML is the fastest growing technical field, at the intersection of informatics and statistics, tightly connected with data science and knowledge

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2016 in
B. Weyers, A. Dittmar (Hrsg.):
Mensch und Computer 2016 – Workshopbeiträge, 4. - 7. September 2016, Aachen.

http://dx.doi.org/10.18420/muc2016-ws11-0001

discovery; and health informatics is among the greatest challenges (Jordan & Mitchell, 2015), (Le Cun, Bengio & Hinton, 2015), (Lake, Salakhutdinov & Tenenbaum, 2015).

In daily life we have often to make decisions without sufficient experience or personal background knowledge of alternatives, consequently we rely on recommendations of other people. Consequently, recommendations are a firm part of natural human social interaction (Taraghi, Grossegger, Ebner & Holzinger, 2013), (Desrosiers & Karypis, 2011).

Apart from these naïve daily observation, decision making/decision support is the supreme discipline in health informatics and the core competency in the health sciences (McNeil, Keeler & Adelstein, 1975), (Croskerry & Nimmo, 2011), (Holzinger, 2014).

Recommender systems primarily assist and augment these natural social processes. In a typical recommender system people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. In some cases, the primary transformation is in the aggregation; in others the system's value lies in its ability to make good matches between the recommenders and those seeking recommendations (Resnick & Varian, 1997).

2 Background and Related Work

The underlying theory of recommender systems is in collaborative filtering. The idea stems from famous Xerox PARC, first applied in the Tapestry system (Goldberg, Nichols, Oki & Terry, 1992).

2.1 Collaborative Filtering

Collaborative filtering (CF) allows users to tag content and have other users benefitting from this tagging. Beyond the manual tagging automatic use-based tagging can be applied. Consequently, CF may be considered a special case of usage mining, which relies on previous recommendations by other users in order to predict which among a set of items are most interesting for the current user (Srivastava, Cooley, Deshpande & Tan, 2000). This helps to answer the question of *"what is interesting?"* (Miller & Sittig, 1990), (Silvia, 2005) which is together with the question "*what is relevant?"* among the grand research questions in decision making and decision support, which is an growing research area in both machine learning and health informatics (Tulabandhula & Rudin, 2014), (Holzinger, 2014).

Systems following such approaches are generally called recommender systems: humans provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. In some cases, the primary transformation is in the aggregation; in others the system's value lies in its ability to make good matches between the recommenders and those seeking recommendations (Resnick & Varian, 1997).

Naturally, recommender systems are in the daily experiences of most Web users today and have manifold applications in different domains, ranging from E-commerce (product

recommendations) to Science (e.g. recommending papers to reviewers) with sheer endless application possibilities. In contexts with increasing information overload, people have to use a variety of strategies to make choices (Jannach, Zanker, Felfernig & Friedrich, 2010).

Technically, recommender systems store a data table that records for each user/item pair whether the user has made a recommendation for the item or not, and also the strength of the recommendation. Typical approaches are content-based or user-based approaches or hybrid approaches. Bayesian classifiers, non-negative matrix factorization or singular value decomposition are used to cluster documents, users and interests for recommendations. These numerical approaches are extended by machine learning approaches, e.g. deep learning approaches (Wang, Wang & Yeung, 2015) or evolutionary computing approaches (da Silva, Camilo, Pascoal & Rosa, 2016).

However, understanding how a neural network completes its task is still hard or impossible to answer. Visual interpretations of neuron-feature relationships have been impressively demonstrated using feed-forward networks in the Deep Dream Project of Google. Using high-dimensional non-visual data, makes this task infinitely more complicated and there are a lot of open research routes for future work.

2.2 Interactive Machine Learning with the human-in-the-loop

Interactive Machine Learning (iML) can be defined as algorithms that can interact with both computational agents and human agents and can optimize their learning behaviour through these interactions (Holzinger, 2016b), (Holzinger, 2016a). In active learning such agents are called oracles (Settles, 2011).

2.3 When is the human-in-the-loop beneficial?

There is evidence that humans sometimes still outperform ML-algorithms, e.g., in the instinctive, often almost instantaneous interpretation of complex patterns, for example, in diagnostic radiologic imaging: A promising technique to fill the semantic gap is to adopt an expert-in-the-loop approach, to integrate the physician's high-level expert knowledge into the retrieval process by acquiring his/her relevance judgments regarding a set of initial retrieval results (Akgul et al., 2011). Despite these apparent findings, so far there is little quantitative evidence on effectiveness and efficiency of iML-algorithms. Moreover, there is practically no evidence, how such interaction may really optimize these algorithms, even though "natural" intelligent agents are present in large numbers on our world and are studied by cognitive scientists for quite a while (Gigerenzer & Gaissmaier, 2011). A very recent work is on building probabilistic kernel machines that encapsulate human support and inductive biases, because state-of-the-art ML algorithms perform badly on a number of extrapolation problems, which otherwise would be very easy to solve for humans (Wilson, Dann, Lucas & Xing, 2015).

2.4 Trust for the doctor-in-the-loop

The doctor-in-the-loop (DiL) as new paradigm in **information driven medicine,** picturing the doctor as authority inside a loop not only supplying an expert system with data and information, but also to interactively manipulate algorithms and tools (Holzinger, 2016a), (Holzinger, 2016b). Before this DiL-paradigm can be implemented in any such system for use in real-world clinical medicine, the trustworthiness of such a system must be assured (O' Donovan & Smyth, 2005). It is well known that publicly accessible adaptive systems such as collaborative recommender systems present a huge security problem, mostly due to the fact that potential attackers cannot easily be distinguished from end users. Apart from technical risks, such attacks may lead to a degradation of user trust in the objectivity and accuracy of such system. A major issue for further research is in modelling attacks and to examine their impact on recommendation algorithms.

One benefit of the DiL paradigm could be that hybrid algorithms may provide a higher degree of robustness (Mobasher, Burke, Bhaumik & Williams, 2007). The doctor as authority inside a loop with an expert system in order to support the (automated) decision making with expert knowledge, not only includes support in pattern finding and supplying external knowledge, but the inclusion of data on actual patients, as well as treatment results and possible additional (side-) effects that relate to previous decisions of semi-automated systems. In this sense, the DiL-concept can be seen as an extension of the increasingly frequent use of knowledge discovery for the enhancement of medical treatments together with human expertise: The expert knowledge of the doctor is enriched with additional information and expert know-how (Kieseberg, Weippl & Holzinger, 2016), (Kieseberg et al., 2016), (Kieseberg, Frühwirt, Weippl & Holzinger, 2015).

3 Future Outlook

We are very much interested in applying recommender systems for solving problems in health informatics, where there is not much previous work. To date as of 5th June 2016, the related work comprises 17 results in the Web of Science with the title “recommender systems health”, the oldest ranging back to 2007 and the most cited having 14 citations:

A five page research statement on the use of recommender systems for personalized health education by (Fernandez-Luque, Karlsen & Vognild, 2009) argues that these systems do not take advantage of the increasing amount of educational resources freely available on the Web, and they point out that it is a difficult problem to find and to match the *relevant* ones.

(Sezgin & Ozkan, 2013) provided at the EHB 2013 a four-page review on health recommender systems, where they emphasize the increasing importance od so-called context, Health Recommender Systems (HRS) which are presented as complementary tools in decision making processes in health care services and have potential to increase usability and acceptance of technologies and reduce information overload in many processes.

A very important future research is in measuring and benchmarking recommender systems, particularly in terms of acceptance of end users (Ziefle, Klack, Wilkowska & Holzinger, 2013) and satisfaction and to personalize the system exactly to the needs, demands and requirements of the end user, and this opens a lot of future research issues, bringing diversity and personalization not just to the contents of recommendation lists, but to the recommendation process itself (Zhou et al., 2010). Quality issues of recommender systems will be crucial for the application in the health domain.

Most of all, more comprehensive **quality measures** are urgently sought, but need much theoretical and experimental future work (Herlocker, Konstan, Terveen & Riedl, 2004). The problem is still that most metrics focus on accuracy and ignore e.g. serendipity and coverage. For answering the question "what is interesting?", which is highly important for health informatics. There are well-known techniques by which algorithms can trade-off reduced serendipity and coverage for improved accuracy (such as only recommending items for which there are many ratings). Since users value all three attributes in many applications, these algorithms may be more accurate, but less useful – for algorithm designers this is a difficult task, where again the DiL-paradigm can be very helpful, because the question "what is interesting?" is inherently subjective and of human nature. We need comprehensive quality measures that combine accuracy with other serendipity and coverage, so algorithm designers can make sensible trade-offs to serve users better.

Serendipity is discovery of interesting items by accident, and is one of the cornerstones of scientific progress. However, "what is interesting" is a hard question, and is even hard to define as it is an essentially human construct (Beale, 2007).

Another very important research issue is **trust** in recommender systems, as they have proven to be an important response to the information overload problem, by providing end users with more proactive and personalized information services in the past (O' Donovan & Smyth, 2005), but there are a lot of open research questions in the factors that play roles in guiding recommendations, and must particularly emphasize gender and age (Ziefle, Röcker & Holzinger, 2011). This is also related to user satisfaction, and (Herlocker, Konstan, Terveen & Riedl, 2004) recommend that four questions deserve future attention: 1) For different metrics, what is the level of change needed before end users notice or user behaviour changes? 2) To which metrics are end users most sensitive? 3) How does end user sensitivity to accuracy depend on other factors such as the interface? 4) How do factors such as coverage and serendipity affect user satisfaction? Moreover, Herlocker et al. (2004) state that if these questions are answered, it may be possible to build a predictive model of user satisfaction that would permit more extensive offline evaluation. By the way, we emphasize always the term end user, intentionally, as it will be of particular importance to focus on particular end user groups, e.g. health practitioners, clinical doctors, biomedical researchers where there will be great differences among them.

Also very interesting is the combination of collaborative filtering with content-based approaches to recommender systems, i.e., approaches that make predictions based on background knowledge of specific characteristics of end users, which is a huge topic in preference learning (Fürnkranz, Hüllermeier, Cheng & Park, 2012).

Today, recommender systems are assisting Web users in the daily process of identifying items that fulfil their wishes, requirements, demands and needs and have been applied in E-commerce settings for quite a while with extreme success – and still needing much future research (Felfernig et al., 2013).

Tomorrow, the next big thing is in the application of such system in the health informatics domain, for the benefit of patients, doctors and hospital managers – just everybody to stay healthy and fit.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments on an earlier version of this manuscript. The authors thank the German Research Council DFG for the friendly support of the research in the excellence cluster „Integrative Production Technology in High Wage Countries“.

References

Akgul, C. B., Rubin, D. L., Napel, S., Beaulieu, C. F., Greenspan, H. & Acar, B. 2011. Content-Based Image Retrieval in Radiology: Current Status and Future Directions. *Journal of Digital Imaging*, 24, (2), 208-222, doi:10.1007/s10278-010-9290-9.

Beale, R. 2007. Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and Web browsing. *International Journal of Human-Computer Studies*, 65, (5), 421-433.

Croskerry, P. & Nimmo, G. 2011. Better clinical decision making and reducing diagnostic error. *The journal of the Royal College of Physicians of Edinburgh*, 41, (2), 155-162.

Da Silva, E. Q., Camilo, C. G., Pascoal, L. M. L. & Rosa, T. C. 2016. An evolutionary approach for combining results of recommender systems techniques based on collaborative filtering. *Expert Systems with Applications*, 53, 204-218, doi:10.1016/j.eswa.2015.12.050.

Desrosiers, C. & Karypis, G. 2011. A comprehensive survey of neighborhood-based recommendation methods. *Recommender Systems Handbook*, 107-144.

Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F. & Reiterer, S. 2013. Toward the next generation of recommender systems: applications and research challenges. *Multimedia Services in Intelligent Environments.* Springer, pp. 81-98.

Fernandez-Luque, L., Karlsen, R. & Vognild, L. K. Challenges and opportunities of using recommender systems for personalized health education. MIE, 2009. 903-907.

Fürnkranz, J., Hüllermeier, E., Cheng, W. & Park, S.-H. 2012. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine Learning*, 89, (1-2), 123-156, doi:10.1007/s10994-012-5313-8.

Gigerenzer, G. & Gaissmaier, W. 2011. Heuristic Decision Making. *Annual Review of Psychology*, 62, 451-482, doi:10.1146/annurev-psych-120709-145346.

Goldberg, D., Nichols, D., Oki, B. M. & Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35, (12), 61-70.

Herlocker, J. L., Konstan, J. A., Terveen, K. & Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems,* 22, (1), 5-53, doi:10.1145/963770.963772.

Holzinger, A. 2014. Lecture 8 Biomedical Decision Making: Reasoning and Decision Support. *Biomedical Informatics.* Springer, pp. 345-377.

Holzinger, A. 2016a. Interactive Machine Learning (iML). *Informatik Spektrum,* 39, (1), 64-68, doi:10.1007/s00287-015-0941-6.

Holzinger, A. 2016b. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? *Springer Brain Informatics (BRIN),* 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

Jannach, D., Zanker, M., Felfernig, A. & Friedrich, G. 2010. *Recommender systems: an introduction*, Cambridge University Press.

Jordan, M. I. & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science,* 349, (6245), 255-260, doi:10.1126/science.aaa8415.

Kieseberg, P., Frühwirt, P., Weippl, E. & Holzinger, A. 2015. Witnesses for the Doctor in the Loop. *In:* Guo, Y., Friston, K., Aldo, F., Hill, S. & Peng, H. (eds.) *Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250.* Cham, Heidelberg, Berlin: Springer, pp. 369-378, doi:10.1007/978-3-319-23344-4_36.

Kieseberg, P., Malle, B., Frühwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. *Brain Informatics,* 1-11, doi:10.1007/s40708-016-0046-2.

Kieseberg, P., Weippl, E. & Holzinger, A. 2016. Trust for the Doctor-in-the-Loop. *European Research Consortium for Informatics and Mathematics (ERCIM) News: Tackling Big Data in the Life Sciences* 104, (1), 32-33.

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science,* 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

Le Cun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *Nature,* 521, (7553), 436-444, doi:10.1038/nature14539.

Mcneil, B. J., Keeler, E. & Adelstein, S. J. 1975. Primer on Certain Elements of Medical Decision Making. *New England Journal of Medicine,* 293, (5), 211-215, doi:doi:10.1056/NEJM197507312930501.

Miller, P. L. & Sittig, D. F. 1990. The Evaluation of Clinical Decision Support Systems - What is Necessary versus What is Interesting. *Medical Informatics,* 15, (3), 185-190.

Mobasher, B., Burke, R., Bhaumik, R. & Williams, C. 2007. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT),* 7, (4), 23, doi:10.1145/1278366.1278372.

O' Donovan, J. & Smyth, B. Trust in recommender systems. Proceedings of the 10th international conference on Intelligent user interfaces (IUI 2005), 2005. ACM, 167-174, doi:10.1145/1040830.1040870.

Resnick, P. & Varian, H. R. 1997. Recommender systems. *Communications of the ACM,* 40, (3), 56-58, doi:10.1145/245108.245121.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development,* 3, (3), 210-229, doi:10.1147/rd.33.0210.

Settles, B. 2011. From theories to queries: Active learning in practice. *In:* Guyon, I., Cawley, G., Dror, G., Lemaire, V. & Statnikov, A. (eds.) *Active Learning and Experimental Design Workshop 2010.* Sardinia: JMLR Proceedings, pp. 1-18.

Sezgin, E. & Ozkan, S. A systematic literature review on Health Recommender Systems. E-Health and Bioengineering Conference (EHB), 2013, 2013. IEEE, 1-4.

Silvia, P. J. 2005. What is interesting? Exploring the appraisal structure of interest. *Emotion,* 5, (1), 89.

Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.-N. 2000. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter,* 1, (2), 12-23.

Taraghi, B., Grossegger, M., Ebner, M. & Holzinger, A. 2013. Web Analytics of user path tracing and a novel algorithm for generating recommendations in Open Journal Systems. *Online Information Review,* 37, (5), 672-691, doi:10.1108/OIR-09-2012-0152.

Tulabandhula, T. & Rudin, C. 2014. On combining machine learning with decision making. *Machine Learning,* 97, (1-2), 33-64, doi:10.1007/s10994-014-5459-7.

Wang, H., Wang, N. & Yeung, D.-Y. Collaborative deep learning for recommender systems. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. ACM, 1235-1244.

Wilson, A. G., Dann, C., Lucas, C. & Xing, E. P. The Human Kernel. *In:* Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R., eds. Advances in Neural Information Processing Systems, NIPS 2015, 2015 Montreal. 2836-2844.

Zhou, T., Kuscsik, Z., Liu, J. G., Medo, M., Wakeling, J. R. & Zhang, Y. C. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences of the United States of America,* 107, (10), 4511-4515, doi:10.1073/pnas.1000488107.

Ziefle, M., Klack, L., Wilkowska, W. & Holzinger, A. 2013. Acceptance of Telemedical Treatments – A Medical Professional Point of View. *In:* Yamamoto, S. (ed.) *Human Interface and the Management of Information. Information and Interaction for Health, Safety, Mobility and Complex Environments, Lecture Notes in Computer Science LNCS 8017.* Berlin Heidelberg: Springer pp. 325-334, doi:10.1007/978-3-642-39215-3_39.

Ziefle, M., Röcker, C. & Holzinger, A. 2011. Medical Technology in Smart Homes: Exploring the User's Perspective on Privacy, Intimacy and Trust. *35th Annual IEEE Computer Software and Applications Conference Workshops COMPSAC 2011.* Munich: IEEE, pp. 410-415, doi:10.1109/COMPSACW.2011.75.

Authors

Holzinger, Andreas

Currently, Andreas is Visiting Professor for Machine Learning in Health Informatics 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 biomedical informatics and the life sciences. Andreas obtained a Ph.D. in Cognitive Science from Graz University in 1998 and his Habilitation in Computer Science from Graz University of Technology in 2003. Andreas is Associate Editor of Knowledge and Information Systems (KAIS), and member of IFIP WG 12.9 Computational Intelligence.

Calero Valdez, André

André Calero Valdez has studied computer science at the RWTH Aachen University and holds a PhD in Psychology also from RWTH Aachen University. He a senior researcher at the Human-Computer Interaction Center of the RWTH Aachen University and visiting professor with the HCI-KDD group in Graz, Austria. His thesis dealt with the topic of user-centered design of small screen devices for diabetes patients. He currently conducts research in the topics of knowledge management, social media, and decision support by visualizations. The aim is to manage complexity of information by applying human-computer interaction principles.

Ziefle, Martina

Martina Ziefle holds the chair of communication science and is founding member of the Human-Computer Interaction Center of the RWTH Aachen University. Her research addresses the communication between human-human and human-machine with the research focus on technology acceptance for various technologies with respect to user diversity.