Maxi MUSTER, BSc

On a crazy scientific question and its experimental engineering proof

Master's Thesis

to achieve the university degree of

Master of Science (MSc)

Master's degree programme:

Software Development and Business Management

Graz University of Technology

Supervisor:
Assoc. Prof. Dr. Andreas HOLZINGER
Institute for Information Systems and Computer Media
Graz University of Technology

Graz, 2nd April 2016

This page intentionally left blank

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, DATE

____

YOUR NAME

This page intentionally left blank

Acknowledgements

This page intentionally left blank

Abstract

Keywords

KEYWORDS

ÖSTAT classification

ÖSTAT CLASSIFICATION

ACM classification

ACM CLASSIFICATION

This page intentionally left blank

Kurzfassung

Schlüsselwörter

KEYWORDS GERMAN

ÖSTAT Klassifikation

ÖSTAT CLASSIFICATION

ACM Klassifikation

ACM CLASSIFICATION

This page intentionally left blank

Table of Contents

Introduction and Motivation for Research

This is just some input test text to check out the Tex-file and is taken from (Holzinger, 2016):

Originally the term “machine learning" 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, machine learning (ML) is the fastest growing technical field, at the intersection of informatics and statistics, tightly connected with data science and knowledge discovery, and health is amongst the greatest challenges (Jordan and Mitchell, 2015), (LeCun, Bengio, and Hinton, 2015).

Particularly, probabilistic ML is extremely useful for health informatics, where most problems involve dealing with uncertainty. The theoretical basis for the probabilistic ML was laid by Thomas Bayes (1701–1761),(Bayes, 1763), (Barnard and Bayes, 1958). Probabilistic inference vastly influenced artificial intelligence and statistical learning and the inverse probability allows to infer unknowns, learn from data and make predictions (Hastie, Tibshirani, and Friedman, 2009),(Murphy, 2012).

The application of ML methods in biomedicine and health can, for instance, lead to more evidence-based decision-making and helping to go towards *personalized medicine* (Holzinger, 2014b).

According to Tom Mitchell (Mitchell, 1997), a scientific field is best defined by the questions it studies: ML seeks to answer the question *“How can we build algorithms that automatically improve through experience, and what are the fundamental laws that govern all learning processes?"*

ML is very broad and deals with the problem of extracting features from data to solve predictive tasks, including decision support, forecasting, ranking, classifying (e.g., in cancer diagnosis), detecting anomalies (e.g., virus mutations) or sentiment analysis The challenge is to discover relevant *structural* patterns and/or *temporal* patterns (“knowledge") in such data, which are often hidden and not accessible to the human expert. The problem is that a majority of the data sets in the biomedical domain are weakly-structured and non-standardized (Holzinger, Dehmer, and Jurisica, 2014), and most data is in dimensions much higher than 3, and despite human experts are excellent in pattern recognition for dimensions ≤ 3, such data make manual analysis often impossible.

Most colleagues from the ML community are concentrating on *automatic* machine learning (aML), with the grand goal of bringing humans-out-of-the-loop, and a best practice real-world example can be found in autonomous vehicles.

However, biomedical data sets are full of uncertainty, incompleteness etc. (Holzinger, 2014a), they can contain missing data, noisy data, dirty data, unwanted data, and most of all, some problems in the medical domain are hard, which makes the application of fully automated approaches difficult or even impossible, or at least the quality of results from automatic approaches might be questionable. Moreover, the complexity of sophisticated machine learning algorithms has detained non-experts from the application of such solutions. Consequently, the integration of the knowledge of a domain expert can sometimes be indispensable and the interaction of a domain expert with the data would greatly enhance the knowledge discovery process pipeline. Hence, *interactive* machine learning (iML) puts the “human-in-the-loop" to enable what neither a human nor a computer could do on their own. This idea is supported by a synergistic combination of methodologies of two areas that offer ideal conditions towards unraveling such problems: Human-Computer Interaction (HCI) and Knowledge Discovery/Data Mining (KDD), with the goal of supporting human intelligence with machine intelligence to discover novel, previously unknown insights into data (HCI-KDD approach (Holzinger, 2013)).

We define iML-approaches as algorithms that can interact with *both computational agents and human agents* *) and can optimize their learning behaviour through these interactions.

*) In Active Learning such agents are referred to as so-called “oracles".

This article is a brief introduction to iML, discussing some challenges and benefits of this approach for health informatics. It starts by motivating the need of a human-in-the-learning-loop and discusses three potential application examples of iML, followed by a very brief overview on the roots of iML in historical sequence: reinforcement learning (1950), preference learning (1987) and active learning (1996).

The overview concludes with discussing three examples of potential future research challenges, relevant for solving problems in the health informatics domain: multi-task learning, transfer learning and multi-agent hybrid systems. The article concludes with emphasizing that successful future research in ML for health informatics, as well as the successful application of ML for solving health informatics problems needs a concerted effort, fostering integrative research between experts ranging from disciplines such as data science to visual analytics. Tackling such complex research undertakings needs both disciplinary excellence and cross-disciplinary networking without boundaries.

The first question we have to answer is: *"What is the difference between the iML-approach to the aML-approach, i.e. unsupervised learning, supervised or semi-supervised learning?"*

Generally, ML can be categorized into two large subfields: unsupervised learning and supervised learning. The goal in supervised learning (aka predictive learning) is to learn a mapping (prediction) from input data $\mathbf{x}$ to output data y, given a (human) labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, where $\mathcal{D}$ is the training set containing a number of training samples, e.g. $\mathbf{x}_i$ can be a D-dimensional vector, called *feature vector*, but it can also be a complex data object (image, graph, time series, etc.). Basically, in supervised learning the value of the outcome data is based on the number of input data. In unsupervised learning (aka descriptive learning), there is no outcome data, and the goal is to describe the associations and patterns among a set of input data, i.e. we have only given inputs $\mathcal{D} = \{\mathbf{x}_i\}$, and the goal is to discover patterns (aka knowledge) in the data. This is a much more difficult problem.

Theoretical Background

Related Work

Current System

Materials and Methods

Results

Discussion and Lessons Learned

Conclusions

Future Work

List of Figures

List of Tables

Bibliography

Barnard, George A and Thomas Bayes (1958). “Studies in the history of probability and statistics: IX. Thomas Bayes’s essay towards solving a problem in the doctrine of chances”. In: *Biometrika* 45.3/4, pp. 293–315. DOI: 10.2307/2333180. URL: http://www.jstor.org/stable/2333180.

Bayes, Thomas (1763). “An Essay towards solving a Problem in the Doctrine of Chances (Posthumous communicated by Richard Price)”. In: *Philosophical Transactions* 53, pp. 370–418.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.* New York: Springer. DOI: 10.1007/978-0-387-84858-7.

Holzinger, Andreas (2013). “Human-–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together?” In: *Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127*. Ed. by Alfredo Cuzzocrea et al. Heidelberg, Berlin, New York: Springer, pp. 319–328. URL: https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=382991&pCurrPk=72064.

Holzinger, Andreas (2014a). *Biomedical Informatics: Discovering Knowledge in Big Data*. New York: Springer. DOI: 10.1007/978-3-319-04528-3. URL: http://dx.doi.org/10.1007/978-3-319-04528-3.

Holzinger, Andreas (2014b). “Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning”. In: *IEEE Intelligent Informatics Bulletin* 15.1, pp. 6–14. URL: http://www.comp.hkbu.edu.hk/~cib/2014/Dec/article2/iib_vol15no1_article2.pdf.

Holzinger, Andreas (2016). “Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop?” In: *Springer Brain Informatics (BRIN)* 3, pp. 1–13. DOI: 10.1007/s40708-016-0042-6. URL: http://dx.doi.org/10.1007/s40708-016-0042-6.

Holzinger, Andreas, Matthias Dehmer, and Igor Jurisica (2014). “Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions”. In: *BMC Bioinformatics* 15.S6, p. I1. DOI: doi:10.1186/1471-2105-15-S6-I1. URL: http://www.biomedcentral.com/1471-2105/15/S6/I1.

Jordan, M. I. and T. M. Mitchell (2015). “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245, pp. 255–260. DOI: 10.1126/science.aaa8415. URL: http://www.sciencemag.org/content/349/6245/255.abstract.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539.

Mitchell, Tom M (1997). *Machine learning*. New York: McGraw Hill.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. Cambridge (MA): MIT press. URL: http://www.cs.ubc.ca/~murphyk/MLbook/index.html.

Samuel, Arthur L (1959). “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3, pp. 210–229.