Resolving Limits Faced by Classical Machine Learning Approaches: Areas of Application for Collaborative Interactive Learning Techniques

Christoph Sonntag[1]
[1]*Intelligent Systems Group, University of Passau, Innstrasse 70, Passau, Germany*
sonntagc@fim.uni-passau.de

***Abstract*—The area of *Machine Learning (ML)* has experienced a high level of research interest in the last few years with it's underlying theory reaching back far into the 18Th century. Due to minimal costs for computation and the massive availability of data in the Internet recent research interest has been mainly focused on *automatic ML (aML)*. In this ML paradigm a model (e.g. classifier, ...) is being trained with a pre-labeled training set in order to make predictions on unknown unlabeled data. Problems arise in domains where data is rarely available or where the labeling process is too expensive (computationally or economically). Furthermore, real-world data can be uncertain, incomplete or contain noisy data. In these cases full automation is not reliable enough or even infeasible for specific domains. However, robust and trustworthy results are mandatory in a lot of applications like health informatics or autonomous vehicles. Therefore, systems are needed that interactively integrate knowledge from various experts, which can be collaborating humans or machines, in order to continuously improve results over a systems whole lifetime: So called *Collaborative Interactive Learning (CIL)* systems. This paper contributes in presenting an overview of classical ML approaches in comparison with CIL and outlines possible areas of application. Abstract kuerzen, 18th Jahrhundert interessiert niemanden.**

1. Introduction

As with most concepts, there is no canonical definition for the term *Machine Learning* but at its most basic form it can be described as algorithms that learn from a given set of training examples $\{(x_i, y_i)\}$ of inputs x_i and outputs y_i in order to make predictions on unknown input data. According to Tom Mitchell, a Computer Scientist from Carnegie Mellon University,ML tries to answer how we can build algorithms that automatically improve through experience and what fundamental laws govern all these learning processes [1]. The area of ML in general is a fast-growing discipline at the intersection of statistics and computer science and has experienced massive research interest in the last decades. With its various application possibilities it has also been an interesting branch for economists and entrepreneurs. Since ML scenarios like *supervised*, *unsupervised* and *semi-supervised* learning, which we will cover later, are heavily data-driven, the Internet with its massive amount of data has contributed to the further development of research in the area of fully automated learning algorithms. Today, we can see the results in a variety of applications [2] [3], not limited to the list below.

- *Text classification, Natural Language Processing (NLP)*. Many current mail programs have built-in spam detection which originally used simple regular expressions in order to detect phrases commonly used in spam mails. With state-of-the-art ML techniques it is not only possible for a mail program to query for a number of words but to the pattern how spam mails are constructed and to adapt itself to new types of spam.
- *Speech recognition*. Most commercial applications for speech recognition use ML to train itself to recognize a users speech input. Umformen. Beschreiben, wofr ML genutzt wird.
- *Image recognition, OCR*. Machine Learning methods have also been successfully applied in domains where it's important to extract information from images such as detecting and classifiying objects or recognize handwritten only handwritten? characters. Image recognition and OCR are therefore for example used in medical diagnosis to detect cancer on radiographs, in autonomous vehicles to detect obstacles and to stay on track or in post-offices to sort envelopes with hand-written addresses. Nochmal eine Quellenangabe?
- *Games*. Computer Games ususally offer a Multiplayer mode where a user can interact and play the game even without a human opponent. For games with a small number of possible moves after each turn, and therefore a realtively small game tree conatining all possible moves, the minimax algorithm [4] can be sufficient. It figures out which next move would minimize the worst-case scenario for all subsequent moves. Therefore, it needs to know all possible moves, which can easily be computationally infeasible for games like Chess or AlphaGo where a general move faces between 35 (Chess) and 250 (AlphaGo) possible subsequent game states. In order to still have challenging opponents modern Games are using ML to gain experience by themselves.
- *Search engines, Recommendation systems*. Nearly all

current search machines use ML in one fashion or another. They mainly use these techniques for *(1) User classification*[q?] in order to offer personalized search results and for *(2) Query classification* in order to "understand' a users query and to provide further meaningful information.

Despite the successfull application of ML in a lot of fields, most of these examples need a sufficient amount of labeled training data (x_i with assigned y_i) in order to make correct predictions on or discover structured patterns in unknown data x_j. However, most data sets in the biomedical domain, in robotics or in other fields, where data is collected from sensor systems or other unreliable sources, are often either not available (e.g. rare diseases, borderline cases in road traffic, ...) or contain noisy data, dirty data or unwanted data due to dirty sensors or poor visibility conditions in camera applications. In addition, a machine learning algorithms prediction is solely based on the training data it has seen before. One might argue, that human decisions are also only based on experience they have made since their childhood but humans are often still superior to most algorithms in terms of the instinctive interpretation of complex patterns. Furthermore, they can learn to recognize structural patterns from very few training data.

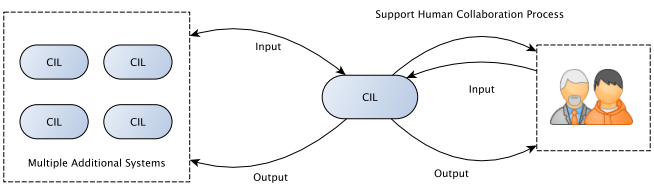

Figure 1. Collaborative Interactive Learning.

Hence, *Collaborative Interactive Learning (CIL)* includes domain experts (humans) and other CIL systems in the decision/prediction making process as shown in Figure 1. The relatively new term describes a new generation of systems with

- lifelong learning capabilities in order to continuously improve its knowledge base
- and the ability to exchange knowledge with other CIL systems as well as humans in order to improve the own knowledge and the one of other entities in a bi-directional way [5].

This article gives a brief introduction on possible areas of application for the CIL-approach. It starts with an overview of classical machine learning scenarios, discussing where they are facing limits and motivating the use of CIL. It then lists possble applications and concludes with a view on future research interests.Relation to Organic Computing?

2. Algorithmic Foundations

Before we take a closer look at the applications and challenges of CIL, first consider the following existing machine learning strategies and algorithms. As already mentioned above, there is a multitude of definitions for the term ML. But primarily, Arthur Samuel coined this term in 1959. Of course, a lot has changed in this area since then, however, the definition that ML allows computers to solve problems without being specifically programmed to do so, is still valid today [9].

Over the last decades, ML has developed significantly and has become an area of research interest for many different actors, which makes it difficult to divide learning strategies into several categories. Different researchers are using different approaches for that [2] [6]. In the following section we want to present five strategies [2].

2.1. Supervised learning

Basically supervised learning is the way of learning from labeled examples and making predictions for all unseen points. This is the most basic concept of ML and used in a variety of applications where labeled data is easy to obtain. The spam mail example discussed in the introduction is an instance of supervised learning.

2.2. Unsupervised learning

In contrast to supervised learning, with this technique the learner only receives unlabeled training data and makes predictions for all unseen points. This technique is relatively similar to the way infants learn new things (e.g. learning a language) and is used, for example, in clustering data into groups [8].

2.3. Semi-supervised learning

As the name suggests, the learner receives training data, which consists of both labeled as well as unlabelled data and makes predictions for all unseen points. This technique is mainly used in areas where data is easily accessible, but finding the right label is expensive.

2.4. Reinforcement Learning

Reinforcement learning is probably the oldest approach and a potentially important one for CIL. The learner interacts with his environment to gain new information. For each action he'll gets a reward. His goal is to maximize his reward over several repetitions. However, he is in conflict with using what has already been won by his actions or to get more information through further actions.

2.5. Active Learning

The goal of Active Learning is to interact with other agents in a clever way in order to label already existing data points and thus achieve a similar performance at less available Information as in the method of supervised learning. To achieve this, the learner needs a selection strategy [7] that decides whether an action has enough information content to label a data point or not.

3. When is using CIL helpful?

In established ML systems, a separate model must be developed and trained in advance for each specific application. ML systems therefore have a relatively narrow application frame. The idea behind CIL systems is that they learn and self-organize knowledge throughout their lives. They collect information from various sources and evaluate the obtained data, so that they can solve problems together with other technical and human agents in a collaborative and interchangeable way [5]. The CIL approach therefore deliberately integrates other intelligent systems (people and machines) into the decision-making process in order to ensure that the individual entities can solve otherwise difficult problems much easier. Another difference to classical approaches is the mutual benefit. Machines not only benefit from training data of other agents, but also provide an own contribution to actively support people's collaboration processes by recognising their wishes and needs and reacting accordingly [5]. D-CIL und O-CIL, obwohl man unterscheidet, betrachten wir hier nur allgemeines Modell.

4. Classification of CIL in Organic Computing

CIL systems can be understood as organically structured information technology which fulfill so-called self-x-properties [10] [**?**]. They

- configure themselves in the sense that they are not aligned by developers to an application case,
- optimize themselves by applying various strategies of ML to expand their knowledge,
- heal and protect themselves because cooperation with other systems is possible.

Depending on which methods of ML are used, CIL systems are also more or less self-explanatory during development [10].

5. Application examples

5.1. Clustering

Clustering is the perfect example of a starting point without sufficient information about a database. The basic goal here is to divide a data set into homogeneous groups of individual data points in order to create structure. Clustering is used, for example, in most recommendation systems of popular Internet applications (Netflix movie recommendations, YouTube Video Tags [?]) to group similar tags or find groups of people with a similar taste in movies. Furthermore, clustering is used to identify communities in social networks, for example for target-advertising, or to cluster location-data so that "important places" of a person can be identified.

Unfortunately, real-world data in these areas is often unclean, contains false or unwanted data and is often present in dimensions >3, so that people alone, who are otherwise very good at recognizing similarities, are dependent on ML and vice-versa.

There are different approaches to cluster data into homogeneous groups. One of them is the k-Means algorithm[?]. This requires at the beginning the number of k clusters it should find. It then randomly places k centroids in space, which should represent the center of a cluster, and assigns each item to the nearest centroid. After the assignment, each centroid is placed at the center of all items assigned to it. This process of assigning the items and moving the centroids to the center is now repeated until the assignments of the items no longer change.

What all cluster algorithms have in common is that they need a way to compare data points for similarity. Classically, data points are tried to display in an n-dimensional metric space and, for example, their Euclidean distance is measured for similarity. The Euclidean distance between two data points $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ and thus their similarity, when a small result means more similar, is defined as

$$similarity(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \quad (1)$$

Since data is usually multidimensional, i.e. much larger than three, suitable attributes must be selected for the comparison. However, the problem here is that similarity functions are influenced by subjective factors, since experts from different fields also find different partial aspects of the amount of data to be investigated interesting. Therefore, the CIL approach can provide a decisive advantage in this case, as the knowledge and assessment of different experts and other systems can be used to make such decisions.

5.2. Health Domain

5.3. Industry and Manufacturing

5.4. Agriculture

- Clustering
  - (clustering communities on Facebook for group target advertising (politics))
  - Methods: k-Means, DJ-clustering? –¿ add paper ref
- Health
  - Detecting cancer on radon, Prof. Sauer/Forwiss research?
  - Growing number of users are using smartphones sensors and related devices for collating a range of health information, track regularly and detect patterns (noise?, privacy?)
- Industry and Manufacturing
  - quality assurance
- Agriculture
  - Detecting bad grains, → Hackzurich project

6. Challenges and future directions

Challenges Ethical Questions

References

[1] Mitchell, T. M. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 45(37), 870-877.

[2] Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.

[3] Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

[4] Bachmaier, C. (2017). Lecture. Programming II. Faculty for Informatics and Mathematics, University of Passau, Germany.check this

[5] Sick, B., Oeste-Reiß, S., Schmidt, A., Tomforde, S., & Zweig, A. K. (2018). Collaborative Interactive Learning. Informatik-Spektrum, 41(1), 52-55.

[6] Corne, D., Dhaenens, C., & Jourdan, L. (2012). Synergies between operations research and data mining: The emerging use of multi-objective approaches. European Journal of Operational Research, 221(3), 469-479.

[7] Calma, A., Leimeister, J. M., Lukowicz, P., Oeste-Reiß, S., Reitmaier, T., Schmidt, A., ... & Zweig, K. A. (2016, April). From active learning to dedicated collaborative interactive learning. In ARCS 2016; 29th International Conference on Architecture of Computing Systems; Proceedings of (pp. 1-8). VDE.

[8] Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. Brain Informatics, 3(2), 119-131.

[9] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

[10] Müller-Schloer, C., von der Malsburg, C., & Würt, R. P. (2004). Organic computing. Informatik-Spektrum, 27(4), 332-336.

[11] Schmeck, H. (2005, May). Organic computing-a new vision for distributed embedded systems. In Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005. Eighth IEEE International Symposium on (pp. 201-203). IEEE.

[12] The Science Behind Foldit. (n.d.). Retrieved June 29, 2018, from https://fold.it/portal/info/about. University of Washington, Departments of Computer Science & Engineering and Biochemistry.