Bulletin of the *Transilvania* University of Braşov • Vol 11(60), No. 2 - 2018
Series III: Mathematics, Informatics, Physics, 273-278

MACHINE LEARNING IN HEALTHCARE: AN OVERVIEW

Árpád KERESTÉLY[1], Lucian Mircea SASU[2] and Marius Sabin TĂBÎRCĂ[3]

Communicated to:
International Conference on Mathematics and Computer Science ,
June 14-16, 2018, Braşov, Romania, 3rd Edition - MACOS 2018

Abstract

Machine learning has been widely used in many domains lately and an increasing trend can be observed. New algorithms have been developed and older ones have been improved or combined to get better results. Healthcare is one of the domains that has seen the benefits of using these new computation methods. Considering that early prototypes of artificially intelligent doctors already exist [2], in the not too distant future we could be greeted by robot nurses or even doctors at medical facilities. This paper aims to present, analyze and discuss some of the latest advancements in machine learning from the healthcare point of view. Important aspects that this paper covers are: recently used machine learning algorithms in healthcare, data available for research purpose and the fields that healthcare extends to. A conclusion based on these aspects is drawn, whether there is a need and possibility for potential further development of machine learning algorithms in healthcare.

2000 *Mathematics Subject Classification:* 68T01, 68T05.
Key words: machine learning, healthcare.

1 Introduction

Machine learning is considered to be the answer to many complex problems for which researchers couldn't find a solution. How come that until now humankind has still not found any cure to heal diseases like cancer or HIV/AIDS? If a cure has not been found,

[1]Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail: arpad.kerestely@unitbv.ro
[2]Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail: lmsasu@unitbv.ro
[3]Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail: marius-sabin.tabirca@unitbv.ro

then how come that not even predictions can be accurately made for some diseases to eventually prevent them? The answer to these questions is hard to tell but for sure it is either that machine learning is not the answer to these problems or that it is still premature to talk about a verdict, because further research needs to be done. Assuming that the second answer is the right one, an overview of the current status in the domain of machine learning and healthcare needs to be done.

A quick investigation in the field of machine learning and healthcare reveals some interesting facts that are not trivial from the start. Namely, that healthcare is a vast domain that comprises not only the study of cancer (or any other disease) prevention or curing, but also the study of anything related to the human body and its wellbeing. Some of the interesting areas that also compose healthcare are the study of emotions, the monitoring of body with mobile devices like smart watches or the analysis of human interactions in the virtual space for example with social networks or search engines. On the other hand, the machine learning term is used to refer to a collection of algorithms that often use probabilistic methods to give computers the ability to improve their performance on a specific task without being explicitly programmed. The collection of machine learning algorithms grows each year, trying to provide a solution to specific problems. The next chapter will present some of the algorithms used in the recent researches regarding healthcare.

2 Summary of recent researches

It is interesting to see that researchers form all around the world are interested in the topic of machine learning related to healthcare. The studied papers are written in different parts of the world, yet they share the same goal, of bringing improvements to healthcare. For example: [7] was done in Brazil, [2] in California, [1] in China. The aforementioned papers also implemented their researches on the data available locally in their region (it's important to note that some diseases can manifest in different ways depending also on where they were observed [1]).

For most of the machine learning algorithms it is a necessity to have training data to work on. The bigger the data the more accurate the solution will be. The data can come from various sources (see Figure 1) and lately the amount has increased quite fast [5].

For most of the data a researcher needs to pay but there are quite a few publicly available data as well, sign that the community supporting this progress is growing. More public data means that researchers who don't have access to a medical facility's data, can also start researching. It is important to note from Figure 1, that healthcare data come from medical facilities in the form of medical records, but it can also come from other sources such as genomic data, internet usage or mobile data. Based on internet usage, although it failed, Google Trends tried to predict epidemic outbreaks between 2008 and 2013. Mobile data can refer to any mobile device attached to the human body (ex. smart watch, smart clothing, electrocardiogram (EKG) signals provider, etc.), and it is growing fast with the development of the IoT field. Some mobile devices can do on-line evaluations, at the current time only simple ones, but as [3] suggested, optimizations can be done to machine learning algorithms so that these devices could ultimately perform a

more powerful processing while not running out of battery in a matter of minutes.

With more and more data available the only issue is to find the appropriate algorithms that can extract useful information. Recent researches proved that machine learning algorithms are capable of such things, although they are not perfect, in many cases they outperform human capabilities or simply provide a starting point for further research.

A system that can be used by medical assistants to classify patients on different Surveillance Levels (SLs) is presented in [7]. A severe SL would mean that immediate medical attention is needed. This classification was done with multiple machine learning algorithms: Artificial Neural Network (ANN), Relevance Feedback (RF), K-nearest Neighbor (KNN), Decision Tree (DT) and Vote (a combination of the aforementioned ones). The authors claim that the accuracy of their system is 87.81%.

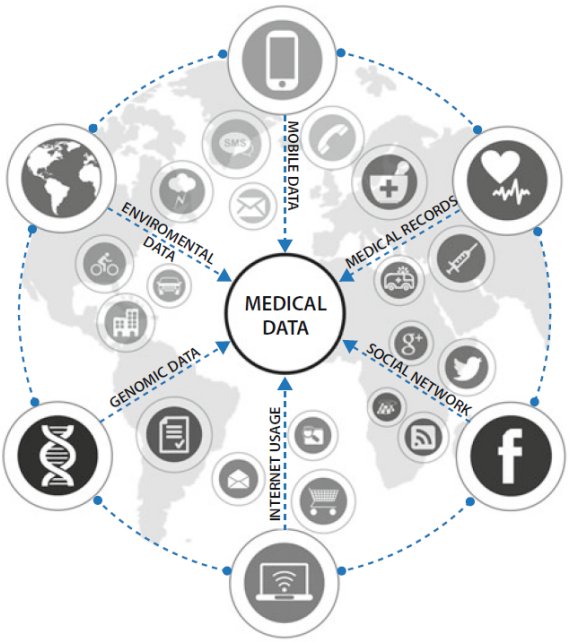

Figure 1: Where does data come from in healthcare [5].

Another paper, namely [1] presented a learning algorithm, that can identify people with chronicle diseases, namely cerebral infarction. The data which they used as input is composed of structured (age, sex, smoking, etc.) and unstructured (text record of patients and/or doctors remarks) data. The algorithm used is CNN-MDRP, Convolutional Neural Network-based multimodal disease risk prediction. This algorithm is a combination of CNN-based unimodal disease risk prediction (used to extract features from the unstructured data) and Naïve Bayesian (NB), K-nearest Neighbor (KNN), Decision Tree (DT) (used to extract features from structured data). The accuracy of the algorithm is claimed to be 94.8%.

An algorithm that can handle historical data is presented in [2]. The authors claim that their algorithm can predict a patients next diagnosis, medication and visit time, based on that persons previous records, using the knowledge of all the patients. The algorithm used is a Recurrent Neural Network (RNN). The claimed accuracy is 78% which is not so good compared to the results from the other papers mentioned but better than the accuracy of an average doctor (the authors claim).

As mentioned in the introduction, healthcare extends also to neuroscience. [4] presented a method to determine patients with suicidal risks using machine learning algorithms. A Gaussian Naïve Bayes (GNB) classifier trained on the data of 33 out of 34 participants, predicted the group membership of the remaining participant with an accuracy of 91%. The data was functional magnetic resonance imaging (fMRI) signatures gathered from patients during a session in which different strongly discriminated concepts

like "death", "cruelty", "trouble", "carefree", "good" and "praise", were presented to the patients. The accuracy of the learning algorithm was pretty good considering that the dataset was relatively small.

The recent researches focus not only on finding solutions for current health related problems, but also on protecting the actual and future development of machine learning algorithms. [6] presents methods which could be used by malicious entities to degrade the performance of machine learning algorithms. The author splits the "attacks" into two categories: exploratory, in which the attacker has access to the dataset used for the training of the machine learning algorithm and causative in which the attacker does not have access to the dataset. The focus is mainly on the poisoning attacks which are a subcategory of the causative attacks. Countermeasures are also presented for these kinds of attacks. Notable is that these attacks can affect any machine learning algorithm, although some are more resistant than others. Support Vector Machine (SVM) being the most resistant while BFTree the most vulnerable in the author's experiments.

The aforementioned researches all used some kind of data filtering. While some used manual data filtering [7], examples of semi-automatic [1] or fully-automatic [3] (yet not so efficient) filtering methods were also presented. For validating their results, most papers used the cross validation method.

3 Conclusions

Better and numerous, results and solutions in healthcare are needed that is certain. Data amount has increased recently and it will probably increase even further in the future. Machine learning algorithms have been researched to solve some specific problems, but there is still plenty of space for improvement, as very few of them address problems with high accuracy and even fewer in a generic way. None of the machine learning algorithms used are suitable for more than a few problems; as seen in the previous section, none of them appeared more than twice, which is a sign that probably the better algorithms are still undiscovered. Yet a combination of existing machine learning algorithms proved to be also effective enough.

Considering these facts, the future of healthcare looks promising with the aid of machine learning and although autonomous AI doctors are unlikely to be invented in the near future, medical assistants could be the first step towards healthcare revolution.

References

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, *Disease prediction by machine learning over Big Data from healthcare communities*, IEEE Access, **5** (2017), 8869-8879.

[2] Choi, E., Bahadori, M., Schuetz, A., Stewart, W., Sun, J., *Doctor AI: predicting clinical events via recurrent neural networks*, arXiv:1511.05942 (2016).

[3] Maharatna, K., Bonfiglio, S., *Machine Learning Techniques for Remote Healthcare*, Springer, 2014.

[4] Just, M., Pan, L., Cherkassky, V., McMakin, D., Cha, C., Nock, M., Brent D., *Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth*, Nature Human Behaviour **1** (2017) 911-919.

[5] Ghassemi, M., Celi, L., Stone, D. *State of the art eeview: the data revolution in critical care*, Critical Care **19** (1) (2015), article 118.

[6] Mozaffari Kermani, M., Sur-Kolay, S., Raghunathan, A, Jha, N., *Systematic poisoning attacks on and defenses for machine learning in healthcare*, IEEE Journal of Biomedical **19** (6) (2014) 1893-1905.

[7] Pollettini, J., Panico, S., Daneluzzi J., Tinós R., Baranauskas, J., Macedo, A., *Using machine learning classifiers to assist healthcare-related decisions: classification of electronic patient records*, J. Med. Systems, **36** (6) (2012), 3861-3874.