ISSN: 2581-3404 (Online)

International Journal of Innovative Research in Technology and Management, Vol-4, Issue-6, 2020.

Interpretable Machine Learning in Health Care: Survey and Discussions

Ravindra Kumar Ahirwar[1], Prof. Rakesh Kumar Lodhi[2], Prof. Neetesh Gupta[3]

[1]M. Tech Scholar, Department of CSE, TIT College, Bhopal, M.P., (India)

[2]Assistant Professor, Department of CSE, TIT College, Bhopal, M.P., (India)

[3]Professor, Department of CSE, TIT College, Bhopal, M.P., (India)

ABSTRACT

With data becoming increasingly available in recent years, black-box algorithms like boosting methods or neural networks play more important roles in the real world. However, interpretability is a severe need for several areas of applications, like health care or business. Doctors or managers often need to understand how models make predictions, in order to make their final decisions. The deployment of ML systems in complex, real world settings has led to increasing interest in systems optimized not only for expected task performance but also other important criteria such as safety, nondiscrimination, avoiding technical debt or satisfying the right to explanation.

Keywords:- Health care, Machine learning, Supervised learning, Electronic health record.

INRODUCTION

In recent years, machine learning is widely used in several areas from business to health care. For most problems, people such as doctors or managers need to understand how the model predicts to help them make the final decision. Also, as the European Union’s new General Data Protection Regulation described [6], users can ask for an explanation of an algorithmic decision that was made about them. Interpretability is a severe need when building a machine learning model for real-world applications. However, for most datasets, interpretable machine learning methods rarely achieve cutting edge performances comparing to black-box methods such as deep learning.

With the increasing amount of data, machine learning holds the potential to become indispensable for doctors to make decisions. When we try to use more machine learning methods in the health care system, interpretability is really needed. Doctors need to understand how the model predictions so they can trust the model. Also, doctors can provide useful suggestions to improve the machine learning model if they understand how the model predicts. There are already applications using interpretable machine learning methods such as decision trees to make clinical decisions. Doctors usually evaluate one medical condition about a patient at a time, which is similar to a decision tree. For classification problems, there are some real-world datasets in which the number of samples of a given class is much smaller than the number of samples in other classes. This imbalance leads to the so-called "class imbalance" problem, one of the major problems in data mining.

Accuracy and interpretability are two dominant features of successful predictive models. There is a common belief that one has to trade accuracy for interpretability using one of three types of traditional models, identifying a set of rules, case-based reasoning by finding similar patients identifying a list of risk factors. While interpretable, all of these models rely on aggregated features, ignoring the temporal relation among features inherent to HER data. As a consequence, model accuracy is sub-optimal. Latent-variable time-series models, such as account for temporality, but often have limited interpretation due to abstract state variables [3].

Health care is undergoing unprecedented change, and there is a great potential and demand for personalized care strategies. Personalized medicine, also called precision medicine, has previously focused on optimizing therapy to better fit the genetic makeup of the patient or the disease (e.g., the genetic susceptibility of cancer to specific chemotherapy strategies). The availability of EHR data and advances in machine learning create the potential for another type of personalization of healthcare [4].

II INTERPREATTION IN DATA SCIENCE LIFE CYCLE

Interpretation in the data science life cycle on its own, interpretability is a broad, poorly defined concept. Taken to its full generality, to interpret data means to extract information (of some form) from it. The set of methods falling under this umbrella spans everything from designing an initial experiment to visualizing final results. In this overly general form, interpretability is not substantially different from the established concepts of data science and applied statistics. Instead of general interpretability, we focus on the use of interpretations in the context of ML as part of the larger data science life cycle. We define interpretable machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data.

Before discussing interpretation methods, we first place the process of interpretable ML within the broader data-science life cycle. Below fig presents a deliberately general description of this process, intended to capture most data-science problems. What is generally referred to as interpretation largely occurs in the modeling and post hoc analysis stages, with the problem, data and audience providing the context required to choose appropriate methods.

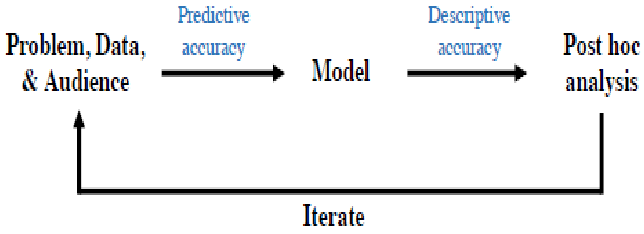

Fig. 1: Overview of different stages (black text) in a data-science life cycle where interpretability is important [2].

III RELATED WORK

[1] In this paper author review the notion of interpretability within the context of healthcare, the various nuances associated with it, challenges related to interpretability which are unique to healthcare and the future of interpretability in healthcare. Applied Machine Learning in Healthcare is an active area of research. The increasingly widespread applicability of machine learning models necessitates the need for explanations to hold machine learning models accountable. While there is not much agreement on the meaning of interpretability in machine learning, there are a number of characteristics of interpretable models that researchers have discussed which can be used as a guide to create the requirements of interpretable models.

[2] Author aim to address these concerns by defining interpretability in the context of machine learning and introducing the Predictive, Descriptive, Relevant (PDR) framework for discussing interpretations. The PDR framework provides three overarching desiderata for evaluation: predictive accuracy, descriptive accuracy and relevancy, with relevancy judged relative to a human audience. Moreover, to help manage the deluge of interpretation methods, we introduce a categorization of existing techniques into model-based and post-hoc categories, with sub-groups including sparsity, modularity and simulatability. To demonstrate how practitioners can use the PDR framework to evaluate and understand interpretations, they provide numerous real-world examples. These examples highlight the

often under-appreciated role played by human audiences in discussions of interpretability. Finally, based on their framework, they discuss limitations of existing methods and directions for future work.

[3] Accuracy and interpretability are two dominant features of successful predictive models. Typically, a choice must be made in favor of complex black box models such as recurrent neural networks (RNN) for accuracy versus less accurate but more interpretable traditional models such as logistic regression. This tradeoff poses challenges in medicine where both accuracy and interpretability are important. They addressed this challenge by developing the reverse Time attentIon model (RETAIN) for application to Electronic Health Records (EHR) data. RETAIN achieves high accuracy while remaining clinically interpretable and is based on a two-level neural attention model that detects influential past visits and significant clinical variables within those visits (e.g. key diagnoses). RETAIN mimics physician practice by attending the EHR data in a reverse time order so that recent clinical visits are likely to receive higher attention.

[4] In this paper author propose a computational framework, Patient2Vec, to learn an interpretable deep representation of longitudinal EHR data, which is personalized for each patient. To evaluate this approach, they apply it to the prediction of future hospitalizations using real EHR data and compare its predictive performance with baseline methods. Patient2Vec produces a vector space with meaningful structure, and it achieves an area under curve around 0:799, outperforming baseline methods. In the end, the learned feature importance can be visualized and interpreted at both the individual and population levels to bring clinical insights.

[6] In this paper they define a ML method as interpretable if the ML model it produces is human-readable, explainable and understandable in domain terms, i.e., is presented in terms of the domain from which the data are used to build said model. An example of such models are those which are produced via decision trees. In a more exact terms, interpretable models are those which are described as statements in propositional or in first order logic terms on predicates from the domain. The proposed DCP algorithm provides a human-friendly visual explanation for its models, in addition to the traditional textual explanation provided via natural language or mathematical forms.

[7] They propose Neural Additive Models (NAMs) which combine some of the expressivity of DNNs with the inherent intelligibility of generalized additive models. NAMs learn a linear combination of neural networks that each attend to a single input feature. These networks are trained jointly and can learn arbitrarily complex relationships between their input feature and the output. Their experiments on regression and classification datasets show that NAMs are more accurate than widely used intelligible models such as logistic regression and shallow decision trees. They perform similarly to existing state-of-the-art generalized additive models in accuracy, but can be more easily applied to real-world problems.

[8] In this paper, they provide an introduction to machine learning tasks that address important problems in genomic medicine. Here, we describe how machine learning can be used to solve key problems in genomic medicine. Genomics is the study of the function and information structure encoded in the DNA sequences of living cells, whereas precision medicine is the practice of tailoring treatment based on all relevant information about the patient, including the patient’s genome.

[9] This article presents a comprehensive up-to-date review of research employing deep learning in health informatics, providing a critical analysis of the relative merit, and potential pitfalls of the technique as well as its future outlook. The paper mainly focuses on key applications of deep learning in the fields of translational bioinformatics, medical imaging, pervasive

sensing, medical informatics, and public health. In this paper, they have outlined how deep learning has enabled the development of more data-driven solutions in health informatics by allowing automatic generation of features that reduce the amount of human intervention in this process. This is advantageous for many problems in health informatics and has eventually supported a great leap forward for unstructured data such as those arising from medical imaging, medical informatics, and bioinformatics.

[12] This article presents a comprehensive overview of the challenges, pipeline, techniques, and future directions for computational health, the rapid growth of novel technologies has led to a significant increase of digital health data in recent years. More medical discoveries and new technologies such as mobile apps, capturing devices, novel sensors, and wearable technology have contributed to additional data sources. Most popular surveys of big data in health informatics have concentrated on biomedical aspects of big data, while a smaller percentage of papers focus on the computational perspective. The ultimate goal of this survey is to connect big data and health informatics communities. The primary emphasis is on computational aspects of big data in health informatics, which includes challenges, current big data mining techniques, strengths and limitations of current works, and an outline of directions for future work.

[14] The purpose of this review is to explore what problems in medicine might benefit from such learning approaches and use examples from the literature to introduce basic concepts in machine learning. It is important to note that seemingly large enough medical data sets and adequate learning algorithms have been available for many decades, and yet, although there are thousands of papers applying machine learning algorithms to medical data, very few have contributed meaningfully to clinical care. The purpose of this review is to explore what problems in medicine might benefit from such learning approaches and use examples from the literature to introduce basic concepts in machine learning. It is important to note that seemingly large enough medical data sets and adequate learning algorithms have been available for many decades, and yet, although there are thousands of papers applying machine learning algorithms to medical data, very few have contributed meaningfully to clinical care. This lack of impact stands in stark contrast to the enormous relevance of machine learning to many other industries.

IV CONCLUSION

The motivation for model explanations in healthcare is clear - in many cases both the end users and the critical nature of the prediction demands a certain transparency both for user engagement and for patient safety. However, merely providing an explanation for an algorithm’s prediction is insufficient. The choice of interpretable models depends upon the application an use case for which explanations are required. Thus a critical application like prediction a patient’s end of life may have much more stringent conditions for explanation fidelity as compared to just predicting costs for a procedure where getting the prediction right is much more important as compared to providing explanations. In this study we present the review of interpretable machine learning for the health care sector.

REFERENCES:

[1] Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, Greg McKelvey, “Interpretable Machine Learning in Healthcare”, IEEE Intelligent Informatics Bulletin, 2018, pp. 1-7.

[2] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-As, Bin Yu, “Interpretable machine learning: definitions, methods, and applications”, IEEE 2018, pp. 1-11.

[3] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, Jimeng Sun, “RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism”, Conference on

Neural Information Processing Systems, 2016, pp. 1-13.

[4] Jinghe Zhang, Kamran Kowsari, James H. Harrison, Jr, Jennifer M. Lobo, Laura E. Barnes, “Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record”, IEEE Access 2018, pp. 65333-65346.

[5] Finale Doshi-Velez, Been Kim, “Towards A Rigorous Science of Interpretable Machine Learning”, 2017, pp. 1-13.

[6] Boris Kovalerchuk, Nathan Neuhaus, “Toward Efficient Automation of Interpretable Machine Learning”, IEEE International Conference on Big Data, 2018, pp. 4933-4940.

[7] Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey E. Hinton, “Neural Additive Models: Interpretable Machine Learning with Neural Nets”, 2020, pp. 1-17.

[8] Michael K. K. Leung, Andrew Delong, Babak Alipanahi, Brendan J. Frey, “Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets”, IEEE Vol-104, 2016. pp. 176-197.

[9] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, Guang-Zhong Yang,” Deep Learning for Health Informatics”, IEEE Journal Of Biomedical And Health Informatics, VOL. 21, 2017. pp. 4-21.

[10] Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, Graciela Gonzalez, “Social Media Mining for Public Health Monitoring and Surveillance”, Pacific Symposium on Bio computing 2016. pp. 468-477.

[11] Zoubin Ghahramani, “Probabilistic machine learning and arti_cial intelligence”, 2015. pp. 1-24.

[12] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, S. S. Iyengar, “Computational Health Informatics in the Big Data Age: A Survey”, ACM Computing Surveys, Vol. 49, 2016. pp. 1-36.

[13] Gunasekaran Manogaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas and Revathi Sundarsekar, “Big Data Knowledge System in Healthcare”, 2017. pp. 133-158.

[14] Rahul C. Deo, “Machine Learning in Medicine”, 2017. pp. 1920-1931.

[15] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T. Dudley, “Deep learning for healthcare: review, opportunities and Challenges”, Briefings in Bioinformatics, 2017, pp. 1–11.

[16] Ji-Jiang Yang, Jianqiang Li, Jacob Mulder, Yongcai Wang, Shi Chen, Hong Wu, Qing Wang, Hui Pan, “Emerging information technologies for enhanced healthcare’, Elsevier ltd. 2015. pp. 3-11.