Kalpa Publications in Computing

Volume XXX, 2022, Pages 8–10

Proceedings of V XoveTIC Conference. XoveTIC 2022

Using Active Learning to Improve the Treatment Selection on Pancreatic Cancer Patients

José Bobes-Bascarán[1], Alberto Pérez-Sánchez[1], Eduardo Mosqueira-Rey[1], David Alonso-Ríos[1], and Elena Hernández-Pereira[1]

Universidade da Coruña (CITIC)

Abstract

The use of Machine Learning (ML) techniques in the context of Cancer prognosis, diagnosis and treatment is nowadays a reality. Some types of cancers could greatly benefit from specific techniques that are designed to work in a scarcity of data scenarios, or when obtaining labeled data is a time-consuming and/or costly task. It is the case of the Pancreatic Adenocarcinoma. We present an experiment where Active Learning (AL) is used as the basis to create a model which performs a classification task where a human expert (in this experiment, a medical doctor) needs to determine whether a pancreatic cancer patient must be treated with chemotherapy, not treated, or he/she is unsure about the therapy. The use of AL techniques allows us to improve the accuracy of the model, and the inclusion of expert opinions may help us in the future to add explanatory capabilities to the system.

1 Introduction

Nowadays Machine Learning (ML) models are being deployed covering a wide range of scenarios, particularly in the health field. Some of the techniques around ML support the creation of models even if the amount of data available is not abundant, or if the annotation process is time consuming. They do so by incorporating humans into the ML loop [4]. This is the case of Pancreatic Cancer and its most common form, Pancreatic Adenocarcinoma. There are few databases providing reliable data about this disease, and in general, the ML models require a big amount of data to be trained in a way that they can offer good results.

Active Learning (AL) [6] is a machine learning technique in which the learner (typically a machine) requests an oracle (typically a human, who acts as a teacher) to label a certain set of examples that could provide relevant information to the learning process, improving its learning performance (i.e., maximizing accuracy) [7, 3]. This technique allows the construction of a model using a small amount of data that is cautiously selected to improve the model performance at each step. It requires the participation of a human domain expert which provides the model with new annotated examples.

In this research, we started from our previous work on AL [1], where we applied some of the ideas in common experimental datasets, and within this work, we did apply those ideas in a real world scenario using specific Pancreatic Cancer data.

2 Gathering and generating data

As mentioned earlier, this study is focused on the Pancreatic Adenocarcinoma, and we use as the initial dataset a publicly available data from The Cancer Genome Atlas (TCGA) [1] that provides over 20.000 characterized primary cancers data, spanning 33 types. For Pancreatic data (project PAAD), there are 185 (102 male and 83 female) cases registered. The database includes patient demographic data, family history, diagnosis, treatments and genomics info. [8].

As the number of cases was scarce, a Generative Adversarial Network (GAN) has been implemented to generate new synthetic cases to be added to the initial set. Particularly, a CTGAN (Conditional GAN for Tabular Data) was used as it allows adding restrictions to the training of the GAN so that it creates new samples as realistic as possible. GANs have been used in the past within medical context to augment datasets with synthetic examples to improve accuracy [2].

3 Active Learning Experiment

In our experiment (see Figure 1), we started from an initial model created by training with a Neural Network over the training set (a subset of the 185 available cases) that we collected from the TCGA-PAAD dataset. Then, combining the non-used cases and the newly generated synthetic cases produced by the GAN, we did develop an Active Learning process to get new labels from a human domain expert (i.e., medical doctor specialized in Oncology). The information was presented using a web application showing one case at a time, requiring the oracle to annotate whether a patient should be given Chemotherapy as a treatment.

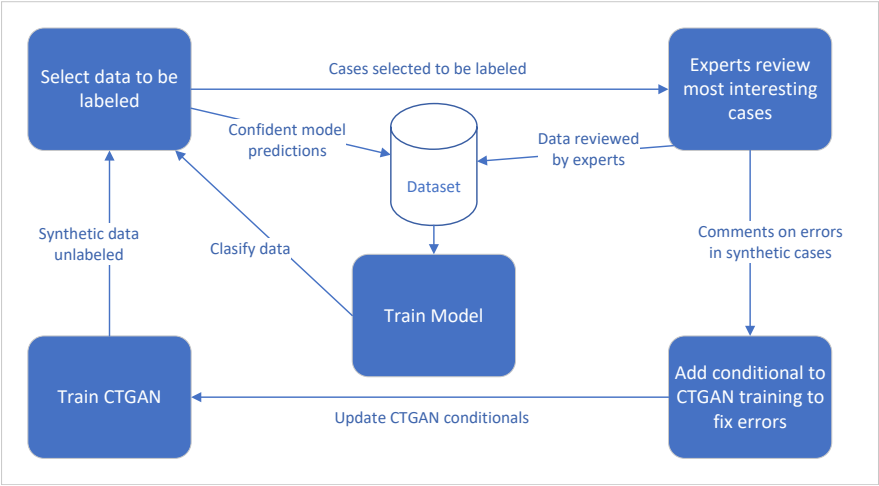

Figure 1: Workflow of the AL experiment.

The experiment was performed using an uncertainty query strategy. This strategy aims at identifying unlabeled elements that are close to the decision boundary and whose class membership is unclear. These elements are most likely to be wrongly classified. Annotations were gathered in a batch of 10 cases, and then fed to training algorithm.

[1] https://www.cancer.gov/tcga

The participation of human experts in the ML loop have provided good results in terms of accuracy increase on each training iteration. Also during the experiment we asked the human experts to add additional information explaining their decisions and analyzing if they consider the case presented as real or synthetic.

4 Conclusions

AL has allowed us to create a reasonable model to determine if chemotherapy should be prescribes on pancreatic cancer patients. This recommendation could serve a medical doctor to help on his/her decision, but it should not be used alone in the clinical set.

Even though the number of examples in the initial model was scarce, the usage of a GAN have provided new useful examples. These examples have been proved to be more realistic on the final sessions, as the medical doctors did have more difficulty distinguishing whether the examples were synthetic or not.

To complete this work, we plan to compare the explanations we have gathered with cancer treatment guidelines [5] analyzing similarities and detecting deviations.

5 Acknowledgments

This work has been supported by Spanish Government (grant PID2019-107194GB-I00 / AEI / 10.13039/501100011033), Xunta de Galicia (grant ED431C 2018/34) and CITIC (grant ED431G 2019/01) with the support of European Union ERDF funds.

References

[1] José Bobes-Bascarán, Eduardo Mosqueira-Rey, and David Alonso-Ríos. Improving medical data annotation including humans in the machine learning loop. *Engineering Proceedings*, 7(1):39, 2021.

[2] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[3] Robert Munro Monarch. *Human-in-the-Loop Machine Learning*. Manning Publications, 2020.

[4] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 2022.

[5] NCCN. *Pancreatic adenocarcinoma, version 3.2019*. National Comprehensive Cancer Network, 2019.

[6] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison. Department of Computer Sciences, 2009.

[7] Burr Settles. From theories to queries: Active learning in practice. In Isabelle Guyon, Gavin Cawley, Gideon Dror, Vincent Lemaire, and Alexander Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy, 16 May 2011. JMLR Workshop and Conference Proceedings.

[8] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):68–77, 2015.