MSc Project Proposal of Optimization of after-care for total hip replacement based on random forest semi-supervised active learning

Chang, Song
Department of Computer Science
University of Exeter
Exeter, UK
cs938@exeter.ac.uk

I. Introduction

Total hip replacement is a highly successful surgery satisficed by patients suffered from hip pain. However, a 20-year assessment for total hip replacement patients shows that there is a 10.7% possibility of revision 20 years after the surgery. [1] Since the spread out of COVID-19, patients who need the beds in hospital has increased.[2] As pressures on the NHS grow, it is increasingly difficult to see patients in person either at the hospital or through a home visit. To maximize the usage of a limited number of beds in the hospital and help doctors to decide which patients are most affected and need to be cared priority, this project will use the combination of semi-supervised learning and active learning to distinguish whether, in which ways and when the patients need follow-up to help to free up these beds.

II. Data & Resources

The data in this project is a unique database of records about the hip replacement operations, after-care, and recovery of patients over 32 years from the Orthopaedical Department at the Royal Devon and Exeter Hospital. The total size of data is which contains roughly 23000 rows and 50 columns. Given that it is a large dataset containing 32 years of records and some patients have done the total hip replacement since the 1990s, we will focus on more recent data for drift and shift of surgical method and material of replaced hip. It is worthily noticed that there is no target label in the dataset, so we need to use active learning which cooperates with doctors to add some labels for training and testing. For protecting the privacy of patients, the data are anonymized and are restored in an encrypted disk.

III. Project Aims and Objectives

A. Does a patient need a follow-up appointment?

As an initial problem, whether a patient needs to be followed-up requires building a classifier to justify the status as ‘necessary’ or ‘not necessary’ by a binary label. Given that the dataset contains limited labels, a semi-supervised classifier [3] will be used for those unlabeled data for its economical and functional features. Random forest classifiers are desirable here, because of their robustness, stability, and ability to handle categorical variables. For those ambiguous classifications, clinicians would check their real categories. Then the checked data will add to the training data. Through this process, the classifier will be retrained, and this loop will continue until a satisfactory accuracy score is obtained. This procedure is also be called active learning. Using these labeled data, a second classifier will be trained to distinguish whether a patient needs to be followed-up. A patient that had no necessary appointments should be classed as not requiring follow-up.

B. Should a follow-up appointment be in-person or via telephone consultation?

After knowing which person required a follow-up appointment, another classifier needs to be trained based on the labels we generated in the previous part. Semi-supervised learning still will be used, because there are limited labeled data for appointment as well. Random forests will also be used here. The ambiguous classification will be distinguished by clinicians and will be labeled. These data then will add to training data set looping until a good accuracy score is got which means active learning will continuously be used. Using this classifier, patients will be classified whether they need an in-person appointment or telephone consultation.

C. When should a patient attend a follow-up appointment?

Since the hospital manager need to arrange the number of beds in each department, it is also important to know how long it will take for the follow-up appointment for each patient. After labeled patients who needs the follow-up, this project will make a prediction for the number of days that the patient needs to attend a follow up appointment. A regression algorithm will be used here. The specific algorithm will depend on the distribution of the dataset or using random forest. In this part, an overall dataset with follow-up

requirements will be used and the regression would predict for both in-person appointment and telephone consultation.

IV. Project Plan

This project is going to start in April. In the beginning, about five weeks is going to be used for data processing such as dealing with missing data and making exploratory data analysis. In this stage, self-implemented decision tree is going to take about 3 weeks and will start from just 2 features for timesaving. After that, twelve weeks will be used for the three problems we have which are whether a patient needs a follow-up, whether they need follow-up through telephone or in-person follow-up and how long should follow-up be taken which would be taken roughly four weeks for each. During the project, we will meet physicians every three weeks to discuss labels and other requirements.

V. Risk Assessment

Because the dataset used is generated in the real world, there would be some limitations probably resulting in the invalidated result. First and foremost, the main risk is that the dataset used for this project doesn't have any labels. Since semi-supervised learning is classified relied on initial labels and that makes it has probability does not work well.

Besides, the dataset contains the records of patients in the last 32 years, but surgical condition or treatment method for surgical aftercare might change from the last 32 years. Since it's historical data, it might be changed in some situations such as copied data to a new system in a system update. There are also some minimum potential risks such as the size of the dataset is the exceptionally large and computational ability for the computer used in the project is limited, the project might face computational risk.

Based on these reasons, this project would concentrate on more recent data as much as possible and that would depend on if there are sufficient ones and try to reduce these risks.

References

[1] L. Neumann, Knude G. F., K. Harry S., Long-term results of Charnley total hip replacement review of 92 patients at 15 to 20 years, March 1994, The Journal of Bone and Joint Surgery, pp. 245-251, available from https://doi.org/10.1302/0301-620X.76B2.8113285

[2] Anita C., Toby Watt, Tim Gardner, Returning NHS waiting times to 18 weeks for routine treatment, May 2020, The Health Foundation, available from https://www.health.org.uk/publications/long-reads/returning-nhs-waiting-times-to-18-weeks

[3] C. Leistner, A. Saffari, J. Santner, and H. Bischof, "Semi-Supervised Random Forests," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 506-513, DOI: 10.1109/ICCV.2009.5459198.

[4] S. Wagner, T. Hastie, B. Efron, “Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife,” Journal of Machine Learning Research, 2014, pp. 1625-1651.