Semi-supervised Graph-based Deep Learning for Multi-Modal Prediction of Osteoarthritis

Alaleh Razmjoo[1], Felix Liu[1], Francesco Caliva[1], Alejandro Morales Martinez[1], Sharmila Majumdar[1], Valentina Pedoia[1]
[1]Center for Intelligent Imaging Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA

Introduction: Osteoarthritis (OA) is a multifactorial disorder accompanied by biochemical and morphological changes in the articular cartilage, modulated by skeletal biomechanics and gait. While we can now acquire detailed information about the knee joint structure and function, we are not yet able to leverage the multifactorial factors for diagnosis and disease management of knee OA. To address this gap, we aim to develop a pipeline for building a semi-supervised predictive model using heterogenous source of data for prediction of OA as expressed by radiographic KL grade. We will use Topological Data Analysis (TDA) to build a multi-domain OA graph to be used as input to a Graph Convolutional Network (GCN) model. This is a novel exploration of the allied used of these tools and it might shed light on the unmet challenge of merging multisource data in deep learning framework, with possible applications beyond the scope of this specific project.

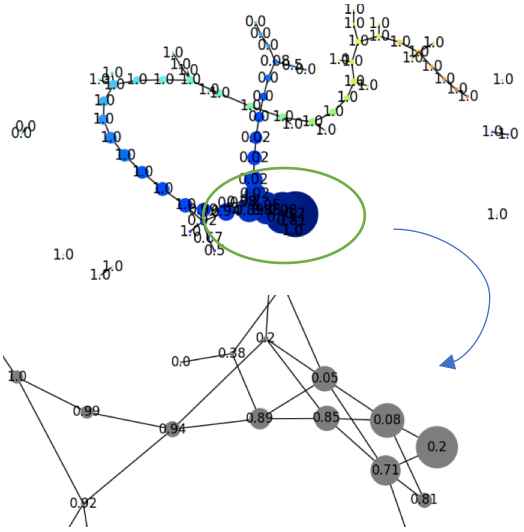

Figure 1: TDA graph of clinical plus shape features. Each node is labeled with proportion of patients with OA (KL >1). The TDA graph is produced using Euclidean distance metric for filter function and single linkage clustering.

Methods: MRI and clinical data is obtained from the Osteoarthrosis Initiative dataset which is a 10 years long study of patients across multiple sites at the baseline timepoint (4469 subjects). Aside from clinical features, the MRI images are used to obtain a dense embedding of bone morphology features from a separately trained deep learning model. To improve the results of the predictive models, feature engineering is performed on the basis of different feature categories. To reduce the number of continuous features, Principal Component Analysis (PCA) is performed to preserve 90% of variance in the feature space. For categorical features, Multiple Correspondence Analysis is performed. The next step in the predictive modeling pipeline is building the patient-graph in which each node contains the data of one patient at baseline. An adapted implementation of TDA mapper tool method for clustering is utilized to build the patient-graph. In a TDA graph, a two-stage clustering operation takes place. First the data points are mapped into a one-dimensional space using a filter function. Then they are binned and clustered based on that filtered value. For example, using eccentricity filter, in a two-dimensional space, data points in the dense areas are considered together for the final clustering. This procedure allows for capturing of topological characteristics of population space. Previous research shows TDA mapper is an effective tool in identifying smaller sub-groups of population that share specific attributes [1]. As stated by [2] the conventional graph building methods such as KNN may not properly identify the sub groups of the population, therefore they used some heuristic approach for graph building which is specific to their problem and requires manual tuning. In this study the graph convolutional neural network architecture (GCN) [3] is used for node labeling. GCNs, are specific type of non-Euclidean neural network designed to work directly on graphs and leverage their structural information. When GCN is used for node labeling it uses single patients features as well as the features of the neighboring nodes in the graph to perform the prediction; in our cases to predict OA using clinical and MRI biomarkers simultaneously. Our semi-supervised graph method was compared with classical ML supervised methods as Random Forest, Multi-Layers Perceptron and Ridge Regression.

Results: The early results show that semi-supervised methods could achieve higher accuracy compared to supervised models when trained with less number of labeled data. Using 5% of data for training TDA + GCN has achieved 83% test accuracy for prediction of OA compared to 79%, 77%, 71% for MLP, ridge classifier, and random forest classifier, respectively. This is a significant result and proves the method could be further applied to the predictive modeling problems where there is limited amount of labeled data for training, as well as the problems with highly imbalanced datasets. Further experiments of this study include the addition of bio-compositional features such as T_2 which are previously shown to be associated with OA. Also fine-tuning of the TDA graph is to be performed to possibly improve the results.

Conclusion: In this study we examined the application of semi-supervised deep learning models for OA prediction using fusion of various feature types. The early results show that TDA could be a viable solution for graph building of graph-based data analysis for disease prediction.

Highlights: We developed a predictive model for OA prediction using the semi-supervised graph-based node labeling model, GCN. The graph building procedure is a novel adaptation of TDA clustering algorithm which is previously used for identification of patients' subgroups. This method is an alternative for fully supervised classification methods which require a lot of labeled data.

References

[1] Nicolau, M., Levine, A.J., and Carlsson, G., Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, Proc Natl Acad Sci U S A. 2011 Apr 26;108(17):7265-70. doi: 10.1073/pnas.1102826108
[2] Parisot S, Ktena SI, Ferrante E, Lee M, Guerrero R, Glocker B, Rueckert D. Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease. Medical image analysis. 2018 Aug 1;48:117-30.
[3] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016 Sep 9.